Bericht zur Datenqualität der GLES: CAPI-Querschnitt 2009, 2013 und 2017 im Vergleich

Bieber, Ina; Etzel, Maximilian

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Bieber, I., & Etzel, M. (2020). *Bericht zur Datenqualität der GLES: CAPI-Querschnitt 2009, 2013 und 2017 im Vergleich.* (GESIS Papers, 2020/12). Köln: GESIS - Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.21241/ssoar.69959

gesis Leibniz-Institut für Sozialwissenschaften

GESIS Papers

2020|12

German Longitudinal Election Study

GLES

Durchgeführt von der Deutschen Gesellschaft für Wahlforschung und GESIS

Bericht zur Datenqualität der GLES: CAPI-Querschnitt 2009, 2013 und 2017 im Vergleich

Ina Bieber & Maximilian Etzel

GESIS Papers 2020|**12**

Bericht zur Datenqualität der GLES: CAPI-Querschnitt 2009, 2013 und 2017 im Vergleich

Ina Bieber & Maximilian Etzel

GESIS – Leibniz-Institut für Sozialwissenschaften 2020

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Dauerbeobachtung der Gesellschaft
GESIS-Projektleitung German Longitudinal Election Study
Postfach 12 21 55
68072 Mannheim

E-Mail: gles@gesis.org

ISSN: 2364-3781 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

1 Einleitung

Querschnittsumfragen bilden den Kern sozialwissenschaftlicher Umfragen weltweit, so auch in der deutschen Wahlforschung. Seit 1949 liegen Querschnittsdaten anlässlich von Bundestagswahlen in Deutschland vor (GESIS, 2019). In Zeiten zunehmender Mobilisierung und Technisierung und auch Pandemien wie COVID 19 stellt sich jedoch die Frage, ob und inwiefern persönlich-mündliche Interviews noch eine geeignete Methode für Sozialforschung im Allgemeinen und die Wahlforschung im Besonderen darstellen, um qualitativ hochwertige Daten zu generieren. Dieser und weiteren Fragen möchte der vorliegende Qualitätsbericht nachgehen.

Es wird die Datenqualität der GLES-Querschnitte, die im Zeitraum von 2009 bis 2017 durchgeführt wurden, auf den Prüfstand gestellt (Rattinger, Roßteutscher, Schmitt-Beck, Weßels, Wagner et al., 2019; Rattinger, Roßteutscher, Schmitt-Beck, Weßels, Wolf et al., 2019; Roßteutscher et al., 2019). Wir fokussieren uns auf diese drei Erhebungen, da der Bericht dazu dienen soll, die aktuelle Qualität der GLES-Querschnitte zu erfassen. Er soll eine angemessene Ausgangsbasis bilden, um entsprechende Strategien zu entwickeln, den GLES-Querschnitt evidenzbasiert in die Zukunft zu führen. Die Fokussierung auf diese drei Zeitpunkte hat zudem den Vorteil, dass die GLES-Querschnitte seit 2009 in der inhaltlichen Ausgestaltung und der methodischen Umsetzung ein sehr homogenes Erhebungsdesign aufweisen, da die Querschnitte als Komponenten der German Longitudinal Election Study (GLES) von 2009 bis 2017 im Rahmen der Langfristförderung der Deutschen Forschungsgemeinschaft gefördert wurden (Bieber & Bytzek, 2013; Schmitt-Beck, Rattinger, Roßteutscher & Weßels, 2010). Seit 2018 wurde die GLES bei GESIS institutionalisiert (GESIS, 2020). Zuvor wurden Wahlstudien von einzelnen Forscher/innen, in Teilen auch mit sehr geringen Ressourcen und Vorlaufzeiten, durchgeführt, was zu divergierenden Frageprogrammen und unterschiedlichen methodischen Standards führte (GESIS, 2019).

Obwohl Datenqualität als ein multidimensionales Konzept zu verstehen ist, fokussiert sich dieser Bericht folgend den bisherigen Datenqualitätsberichten der GLES auf den allgemeinen Aspekt der Repräsentativität, also der Verallgemeinerbarkeit der GLES-Querschnittsdaten auf die interessierte Grundgesamtheit (Roßmann, Blumenberg & Gummer, 2017, S. 12; siehe auch: Blumenberg, Roßmann & Gummer, 2013) und befasst sich darüber hinaus mit dem Feldverlauf und somit der Frage welche Besonderheiten während der Erhebungsphase beobachtet werden konnten. Nachfolgend werden zunächst die Stichprobenqualität und dann die Qualität des Feldverlaufs systematisch dargestellt, analysiert und bewertet.

2 Stichprobenqualität

Eine hohe Qualität der Stichprobenziehung und Stichprobenrealisierung ist für die Durchführung des GLES-Querschnitts zentral, um die kostenintensive Datenerhebung durch persönlich-mündliche Interviews zu legitimieren. Schließlich haben repräsentative Umfragen das Ziel, durch eine angemessene Zufallsstichprobe aus der Grundgesamtheit eine Auswahlgesamtheit zu ziehen, die die Grundgesamtheit bestmöglich abbildet. Eine angemessene Abbildung der Grundgesamtheit in der Auswahlgesamtheit ist für die Verallgemeinerung der gewonnen Erkenntnisse essenziell und Ausdruck qualitativ hochwertiger Umfrageforschung (Bortz & Schuster, 2016; Schumann, 2019). Zwar ist es möglich, durch verschiedenste Verfahren bei und nach der Stichprobenziehung (wie z.B. Registerstichprobe, mehrstufige, geschichtet Auswahl, Gewichtung oder Propensity-Score-Weighting) die Qualität zu optimieren bzw. zu korrigieren (Gabler & Ganninger, 2010; Schnell, 1997), jedoch können auf allen Ebenen der Stichprobenziehung und Stichprobenrealisierung Probleme auftreten, die zu Verzerrungen der Daten und somit zu Verallgemeinerungsproblemen führen (Groves & Lyberg, 2011).

Im Rahmen der GLES-Querschnittsbefragungen wurden 2009 und 2013 Adress-Random-Stichproben und 2017 eine Registerstichprobe gezogen (Rattinger, Roßteutscher, Schmitt-Beck, Weßels, Wagner et al., 2019; Rattinger, Roßteutscher, Schmitt-Beck, Weßels, Wolf et al., 2019; Roßteutscher et al., 2019). Die Datenqualität derartig gezogener, zufallsbasierten Stichproben – sei es nun per Adress-Random oder Registerstichprobe – mit anschließender CAPI-Befragung wird allgemein als sehr hoch eingestuft – es wird auch von dem „Goldstandard“ der sozialwissenschaftlichen Datenerhebung gesprochen (Leeuw & Berzelak, 2016). Dennoch ist dieses Vorgehen keine Garantie für die problemlose Verallgemeinerung der Aussagen, da in der Realität nicht mit jedem ausgewählten Element bzw. nicht mit jeder ausgewählten Adresse auch ein Interview durchgeführt werden kann. Es können sowohl stichprobenneutrale/unsystematische Ausfälle als auch systematische Ausfälle auftreten (Proner, 2011).

Nachfolgend wird die Stichprobenqualität der GLES-Querschnitte anhand von zwei Dimensionen untersucht, wobei sich die erste Dimension dem grundsätzlichen Teilnahmeverhalten der Befragten widmet – also der Frage, welche Personengruppen überhaupt an der Umfrage teilgenommen haben und welche nicht. Die zweite Dimension widmet sich dem inhaltlichen Antwortverhalten der Befragungsteilnehmer/innen und vergleicht dieses mit Auswertungen des Mikrozensus und tatsächlich getätigtem Wahlverhalten. Im Mittelpunkt steht die Frage, ob sich die Personen, die an der Befragung teilgenommen haben, in deutlichem Maße von der eigentlich fokussierten Grundgesamtheit unterscheiden und inwiefern longitudinale Unterschiede zu beobachten sind.

2.1 Teilnahmeverhalten

Das Teilnahmeverhalten der Befragten wird anhand vier gängiger Indikatoren untersucht. Datenbasis hierfür bilden sogenannte Bruttodaten, die mit Informationen aus Kontaktprotokollen bereichert werden. Bruttodaten beinhalten Informationen über alle Personen, die sich in der Stichprobe befinden und mit denen folglich die Interviewer/innen grundsätzlich ein Interview durchführen hätten können. Die Interviewer/innen der GLES-Querschnitte mussten ihre Kontaktversuche mittels sogenannter Kontaktprotokolle dokumentieren. Hier werden die Interviewer/innen verpflichtet, direkt bei der Bearbeitung einer Adresse das Resultat ihrer Arbeit zu notieren: Konnten sie das Interview durchführen? Hat die Zielperson das Interview verweigert? Liegt ein anderer Ausfallgrund vor und soll ein erneuter Kontaktversuch erfolgen? Am Ende der Feldarbeit kann für jede Adresse der Bruttostichprobe nachvollzogen werden, ob, wann und wie häufig ein Kontaktversuch

durchgeführt wurde, ob dieser zu einem Interview geführt hat oder welche Ausfallgründe seitens der Interviewer/innen identifiziert werden konnten. Mittels dieser Rückmeldungen können verschiedene Indikatoren berechnet werden, die Aussagen über die Feldarbeit und das Teilnahmeverhalten ermöglichen.

Zur Standards in der Umfrageforschung haben sich in der sozialwissenschaftlichen Forschung die „Standards der American Association of Public Opinion Research“ (AAPOR) etabliert, die nicht nur modusspezifische Zuteilungen von Ausfallgründen vornehmen, sondern auch verschiedene Berechnungsverfahren anbieten, um das Teilnahmeverhalten einheitlich zu berechnen (AAPOR, 2016). Dies hat den Vorteil, dass das Teilnahmeverhalten international vergleichbar ist und auf definierte Berechnungsverfahren zurückgeführt werden kann.

In diesem Kapitel wird zunächst das allgemeine Teilnahmeverhalten – gemessen an der Ausschöpfung, der Kontaktrate, der Kooperationsrate und der Verweigerungsrate – für die GLES-Querschnitte 2009, 2013 und 2017 dargestellt. Im zweiten Teil wird die Ausschöpfungsquote detaillierter anhand ihrer unterschiedlichen Ausprägung in verschiedenen sozialstrukturellen Gruppen betrachtet. Letzteres kann darüber Aufschluss geben, inwieweit die Befragungsteilnehmer/innen repräsentativ für die Grundgesamtheit stehen.

2.1.1 Ausschöpfung, Kontaktrate, Kooperationsrate und Verweigerungsrate

Die nachfolgende Tabelle 1 listet für die GLES-Querschnittsbefragungen 2009, 2013 und 2017 die erhobenen Ausfallgründe differenziert nach den AAPOR-Standards auf (AAPOR, 2016). Die Zusammenarbeit mit unterschiedlichen Befragungsinstituten 2009 (Marplan), 2013 (Forsa Marplan, nach Insolvenz 2010 als Tochterunternehmen von Forsa aufkauft) und 2017 (KANTAR/TNS Infratest) hat dazu geführt, dass die Ausfallgründe in den drei Befragungen unterschiedlich erhoben wurden. Zudem haben die deutschen Ausfallgründe von persönlich-mündlichen Interviews – trotz internationaler Standardisierung – Besonderheiten in ihrer Erhebung (vgl. hierzu Stadtmüller et al. 2019).

Bei der Zuteilung der Ausfallgründe zu Ausfallgründen nach AAPOR-Standards wurde auf Vorarbeit von Roßmann et al. (2017) zurückgegriffen und eine Spezifizierung von Stadtmüller et al. (2019) zur Zuteilung von Ausfallgründen in deutschen Studien herangezogen. Basierend auf diesen Zuteilungen wurden die verschiedenen Indikatoren berechnet. AAPOR stellt verschiedene Berechnungsverfahren zur Verfügung, wobei nachfolgend die 1. Variante („Contact Rate 1“, „Refusal Rate 1“, „Cooperation Rate 1“, „Response Rate 1“) verwendet wird, die gleichzeitig auch die konservativste Zuteilung vornimmt und im Vergleich zu den anderen Berechnungsverfahren zu niedrigeren Raten gelangt (AAPOR, 2016). Die Formeln lauten folgendermaßen:

Contact Rate 1: (I + P) + R + O / (I + P) + R + O + NC + (UH + UO)

Refusal Rate 1: R / ((I + P) + (R + NC + O) + UH + UO))

Cooperation Rate 1: I / (I + P) + R + O)

Response Rate 1: I / (I + P) + (R + NC + O) + (UH + UO)

I=Complete Interviews (AAPOR-Code: 1.1);
P=Partial Interviews (AAPOR-Code: 1.2);
R=Refusal and break off (AAPOR-Code: 2.1);
NC=No-Contact (AAPOR-Code: 2.2),
O=Other (AAPOR-Codes: 2.0, 2.3);
UH=Unknown Household (AAPOR-Code: 3.1);
UO=Unknown other (AAPOR-Code: 3.2-3.9).

Die Ergebnisse der Berechnungen der verschiedenen Teilnahmeraten sind auch in Abbildung 1 grafisch dargestellt. Die Kontaktrate („Contact Rate“) liegt bei allen Befragungen über 72,0 Prozent

und erreicht ihren Höchstwert bei der Nachwahlbefragung 2017 mit 81,0 Prozent. Diese Zahl gibt den Anteil aller Fälle an, bei denen ein Kontakt mit einem Haushaltsmitglied – sei es die befragte Person oder ein anderes Mitglied – hergestellt werden konnte (AAPOR, 2016, S. 64). Somit zeigt sich, dass mit knapp drei Viertel der Haushalte zumindest ein Kontaktversuch gelungen ist und dass 2017 im Vergleich zu 2013 etwas bessere Raten verzeichnet werden konnten. Ein Abwärtstrend bei der Erreichung der Haushalte kann somit nicht beobachtet werden, wobei hier nicht final geklärt werden kann, worauf die Erhöhung 2017 zurückzuführen ist. Denkbar ist, dass die Feldarbeit oder das Interviewerfeld seitens Marplan, Forsa Marplan und KANTAR variieren und entsprechende Ergebnisse begründen können.

Die Verweigerungsrate („Refusal Rate“) gibt den Anteil der Teilnahmeverweigerungen und Interviewabbrüche an allen teilnahmeberechtigten Fällen an (AAPOR, 2016, S. 64). Es zeigt sich hier deutlich, dass von 2009 zu 2013 die Verweigerungsraten stark gestiegen sind. Zwischen 2013 und 2017 sind geringfügige Änderungen zu beobachten. Lag die Verweigerungsrate 2009 noch unter 30 Prozent (23,1% bzw. 28,1%), erreichte sie 2017 Werte über 40 Prozent (Vorwahl: 40,5%, Nachwahl: 44,1%). Dieser Trend zeigt die zunehmende Schwierigkeit des GLES-Querschnitts, ein Interview mit der ausgewählten Person durchzuführen. Über 40 Prozent sind überhaupt nicht bereit an einem Interview teilzunehmen.

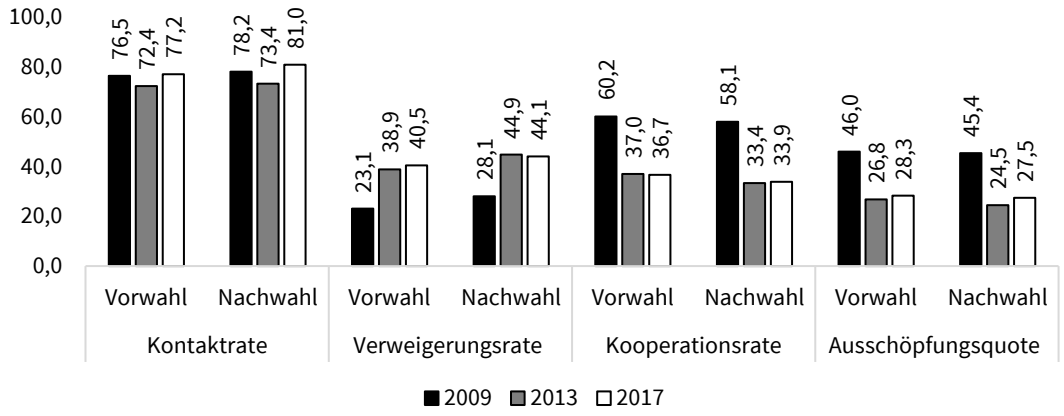


Abbildung 1: Teilnehmer/innenraten der GLES-Querschnitte differenziert nach Vor- und Nachwahlbefragung und Befragungsjahr (in%)

Die Kooperationsrate gibt den Anteil an interviewten Personen an denjenigen Fällen an, die jemals kontaktiert wurden (AAPOR, 2016, S. 63). Die Kooperationsrate wird im vorliegenden Fall als Anteil an durchgeführten Interviews geteilt durch die Summe aus durchgeführten Interviews, Verweigerungen sowie andere Ausfallgründe (Verstorbene, Verzogene, Kranke, Personen mit Sprachproblemen oder unbekannte Ausfälle) berechnet. Auch hier zeigt sich, dass 2009 noch deutlich höhere Kooperationsraten erzielt werden konnten (über 58 Prozent) als 2013 und 2017. 2017 waren nur 36,7 Prozent in der Vorwahl- und 33,9 Prozent in der Nachwahlbefragung zu kooperativem Verhalten bereit. Auffällig ist, dass die Kooperationsraten in der Vorwahlerhebung jeweils höher ausfallen als in der Nachwahlerhebung, was möglicherweise auf das Ereignis der Bundestagswahl selbst zurückgeführt werden kann.

Tabelle 1: Ausfallgründe in den Vor- und Nachwahlbefragungen des GLES-Querschnitts 2009, 2013 und 2017 und berechnete Teilnahmequoten nach AAPOR-Standards

AAPOR-Code	Ausfallgründe	2009				2013				2017			
		Vorwahl		Nachwahl		Vorwahl		Nachwahl		Vorwahl		Nachwahl	
		n	%	n	%	n	%	n	%	n	%	n	%
"Interview"													
1.000	Realisierte Interviews	2173	45,8	2117	45,4	2003	26,2	1908	24,0	2179	28,0	2112	27,2
"Eligible, non-interview"													
2.111	Haushalt verweigert jede Auskunft	720	15,2	845	18,1	1012	13,2	1289	16,2				
2.112	Zielperson verweigert Auskunft	371	7,8	463	9,9	1893	24,7	2204	27,7	3116	40,1	3392	43,6
2.200	Zielperson nicht erreicht	102	2,2	56	1,2	64	0,8	27	0,3	240	3,1	230	3,0
2.240	Im Haushalt niemand angetroffen (Registerstichprobe)									835	10,7	786	10,1
2.250	Kein Termin innerhalb Feldzeit möglich /Zielperson in Feldzeit nicht erreicht	90	1,9	67	1,4	122	1,6	140	1,8	496	6,4	243	3,1
2.310	Zielperson verstorben									30	0,4	50	0,6
2.320	Zielperson dauerhaft krank oder nicht in der Lage dem Interview zu folgen	44	0,9	40	0,9	109	1,4	126	1,6	368	4,7	368	4,7
2.330	Keine Verständigung mögl. (Sprache)	44	0,9	43	0,9								
2.331	Keine Verständigung mögl. (Sprache) - Haushalt					98	1,3	54	0,7				
2.332	Keine Verständigung mögl. (Sprache) - Zielperson					59	0,8	49	0,6	44	0,6	55	0,7
2.360	Sonstige Ausfallgründe	259	5,5	138	3,0	233	3,0	83	1,0	6	0,1	9	0,1
2.410	Zielperson verzogen - innerhalb Deutschlands									123	1,6	168	2,2
2.420	Zielperson verzogen - ins Ausland									16	0,2	22	0,3
2.430	Zielperson lebt in Anstalt									52	0,7	56	0,7
"Unknown eligibility, non-interview"													
3.110	Eingesetzte, vom Interviewer nicht bearbeitete Adressen	527	11,1	498	10,7	1014	13,3	817	10,3				
3.170	Nicht möglich das Haus zu erreichen, zu gefährlich					2	0,0	10	0,1				
3.180	Lokalisierung der Adresse nicht mögl.			2	0,0	53	0,7	40	0,5				
3.200	Im Haushalt niemand angetroffen (Address Random)	389	8,2	392	8,4	804	10,5	1040	13,1				
3.810	Zielperson verzogen - unbekannt									185	2,4	201	2,6
"Not eligible"													
4.190	Adresse falsch, existiert nicht (mehr)									78	1,0	84	1,1
4.500	Keine Wohnadresse	11	0,2	2	0,0	24	0,3	4	0,1				
4.600	Adresse nicht bewohnt	10	0,2	5	0,1	64	0,8	67	0,8				
4.700	Keine zur Grundgesamtheit gehörige Person im Haushalt					96	1,3	107	1,3				
4.900	Anderer Grund (keine Rückmeldung durch Interviewer)									6	0,1		
Total		4740	100,0	4668	100,0	7650	100,0	7965	100,0	7774	100,0	7776	100,0
Teilnahmequoten nach AAPOR-Standards													
Ausschöpfungsquote (Response Rate 1)		46,0		45,4		26,8		24,5		28,3		27,5	
Kooperationsrate (Cooperation Rate 1)		60,2		58,1		37,0		33,4		36,7		33,9	
Verweigerungsrate (Refusal Rate 1)		23,1		28,1		38,9		44,9		40,5		44,1	
Kontaktrate (Contact Rate 1)		76,5		78,2		72,4		73,4		77,2		81,0	

Schließlich zur sogenannten „Ausschöpfung“, auf die die Qualitätsmessung von Umfragen häufig reduziert wird und die im Zentrum vieler Qualitätsbeurteilungen von Umfragen steht (Brick & Williams, 2013; Kreuter, 2013). Die Ausschöpfungsquote zeigt den Anteil an Interviews der infrage kommenden Fällen der Stichprobe (AAPOR, 2016, S. 61). Es zeigt sich, dass diese 2009 noch mit knapp 46,0 Prozent (Vorwahl) bzw. 45,4 Prozent (Nachwahl) vergleichsweise hoch ausfiel. 2013 sank diese auf 26,8 Prozent (Vorwahl) bzw. 24,5 Prozent (Nachwahl). Eine leichte Steigerung zur vorherigen Bundestagswahl konnte 2017 beobachtet werden, wo die Ausschöpfung bei 28,3 Pro-

zent (Vorwahl) bzw. 27,5 Prozent (Nachwahl) lag. Ausschöpfungen in Höhe von ca. 30 Prozent sind ähnlich hoch wie die bei vergleichbar durchgeführten persönlich-mündlich Umfragen. Als Beispiele seien hier die Allgemeinen Bevölkerungsumfrage (ALLBUS) 2018 (32,4 Prozent; ALLBUS, 2019) oder der European Value Survey (28,0 Prozent; Christmann, Gummer, Hähnel & Wolf, 2019) genannt. Ähnlich wie bei der Kontaktrate ist schwer zu beurteilen, worauf die leichte Steigerung der Ausschöpfung 2017 zurückzuführen ist. Doch wie nachfolgender Abschnitt zeigen wird, ist zu empfehlen, nicht allein die Ausschöpfungsquote als Gütekriterium heranzuziehen, sondern diese auch im Kontext zu betrachten.

2.1.2 Ausschöpfung in verschiedenen sozialstrukturellen Gruppen

Die Ausschöpfungsquote wird gerne als Indikator zur Beurteilung der Qualität von Stichproben herangezogen (Brick & Williams, 2013; Kreuter, 2013). Zusätzlich lohnt es sich die Ausschöpfungsquote auch differenziert nach verschiedenen sozialstrukturellen Gruppen zu betrachten. Dies zeigt, wie gut die Stichprobenrealisierung nicht nur allgemein, sondern auch innerhalb verschiedener Bevölkerungsgruppen gelungen ist. Basis der Berechnungen bilden auch hier sogenannte Bruttodaten, also Daten aller in der Stichprobe befindlichen Fälle (vgl. unter 2.1). Aufgrund der Registerstichprobenziehung in der GLES-Querschnittserhebung 2017 ist das Geschlecht, das Geburtsjahr, die Ost-West-Zugehörigkeit, das Bundesland, die Gemeindegrößenklasse und die BIK-Regionen der Bruttostichprobe bekannt. Mit diesen Informationen ist es möglich die Ausschöpfungsquoten differenziert nach sozialstrukturellen Gruppen zu berechnen. Da diese Daten nur für das Jahr 2017 vorliegen, beschränkt sich die nachfolgende Auswertung und Analyse auf die Bundestagswahl 2017.

In der Abbildung 2 ist die Differenz der Ausschöpfungsquote in verschiedenen sozialstrukturellen Gruppen und der durchschnittlichen Ausschöpfungsquote differenziert nach Vorwahl- und Nachwahlerhebung dargestellt. Sie zeigt, mit welchen sozialstrukturellen Gruppen besonders viele bzw. besonders wenige Interviews realisiert werden konnten und somit welche sozialstrukturellen Gruppen in der Stichprobe über- und welche unterrepräsentiert sind.

Es ist zu sehen, dass die Ausfälle nicht über alle sozialstrukturellen Gruppen gleich verteilt sind, was auf Repräsentativitätsprobleme der Stichprobenrealisierung hindeutet. Während der Überhang an männlichen Personen in der Realisierung mit ein bis zwei Prozent noch vergleichsweise gering ausfällt, sind deutlichere Unterschiede in den verschiedenen Altersgruppen zu beobachten: In der realisierten Stichprobe sind die sehr jungen Personen (bis 19 Jahre) sehr stark überrepräsentiert. Hinsichtlich des Ausmaßes ist dies zwar überraschend; jedoch ist das überdurchschnittlich starke Interesse an Politik der sehr jungen Altersgruppe dahingehend bekannt, dass die sehr junge Altersgruppe auch überdurchschnittlich an Wahlen teilnimmt (Kroh & Käppner, 2016). Dennoch darf dieses Ergebnis nicht überinterpretiert werden, da die Fallzahl der Stichprobe gering ist. Leicht unterrepräsentiert sind dagegen die Gruppen der 20- bis 49- und der über 70-Jährigen. Während die Gründe für die Ausfälle bei den 20- bis 49-Jährigen in ihrer Mobilität und Berufstätigkeit zu finden sein könnten, kann die niedrigere Quote der über 70-Jährigen dem erhöhten Alter und damit einhergehenden Erkrankungen geschuldet sein – in beiden Fällen sind dies jedoch Vermutungen, die im Rahmen dieses Berichts weder bestätigt noch weiter analysiert werden können. Festgehalten werden kann jedoch, dass grundsätzlich Altersverzerrungen bei der Stichprobenrealisierung 2017 zu beobachten sind.

Besonders stark fällt die Unter- und Überrepräsentationen in einigen Bundesländern aus: In den ohnehin bevölkerungsstarken Bundesländern Bayern, Brandenburg und Rheinland-Pfalz konnten sowohl in der Vor- als auch in der Nachwahlerhebung überproportional viele Interviews realisiert werden. Menschen aus Baden-Württemberg, Berlin, Mecklenburg-Vorpommern, Nordrhein-

Westfalen, Sachsen und Sachsen-Anhalt sind dagegen unterrepräsentiert. Vor- und nachwahlspezifische Unterschiede sind in den bevölkerungsärmeren Bundesländern Bremen, Hamburg und dem Saarland zu beobachten, ebenso wie in Hessen. Grundsätzliche Probleme in der Interviewrealisierung in den verschiedenen Bundesländern, sowie insbesondere in den Stadtstaaten und im Saarland sind bereits bekannt. Ihnen wurde auch während der Datenerhebungsphase besondere Aufmerksamkeit geschenkt. Trotz alledem sind die dargestellten Abweichungen zu beobachten. Worauf die Unterschiede allerdings im Detail zurückzuführen sind, kann an dieser Stelle final nicht geklärt werden. Mögliche Erklärungsfaktoren können Probleme mit Interviewer/innen, Probleme im Feldmanagement oder sonstige regionale Faktoren, wie zum Beispiel Bevölkerungsdichte, sozialstrukturelle Faktoren oder stichprobenspezifische Besonderheiten sein.

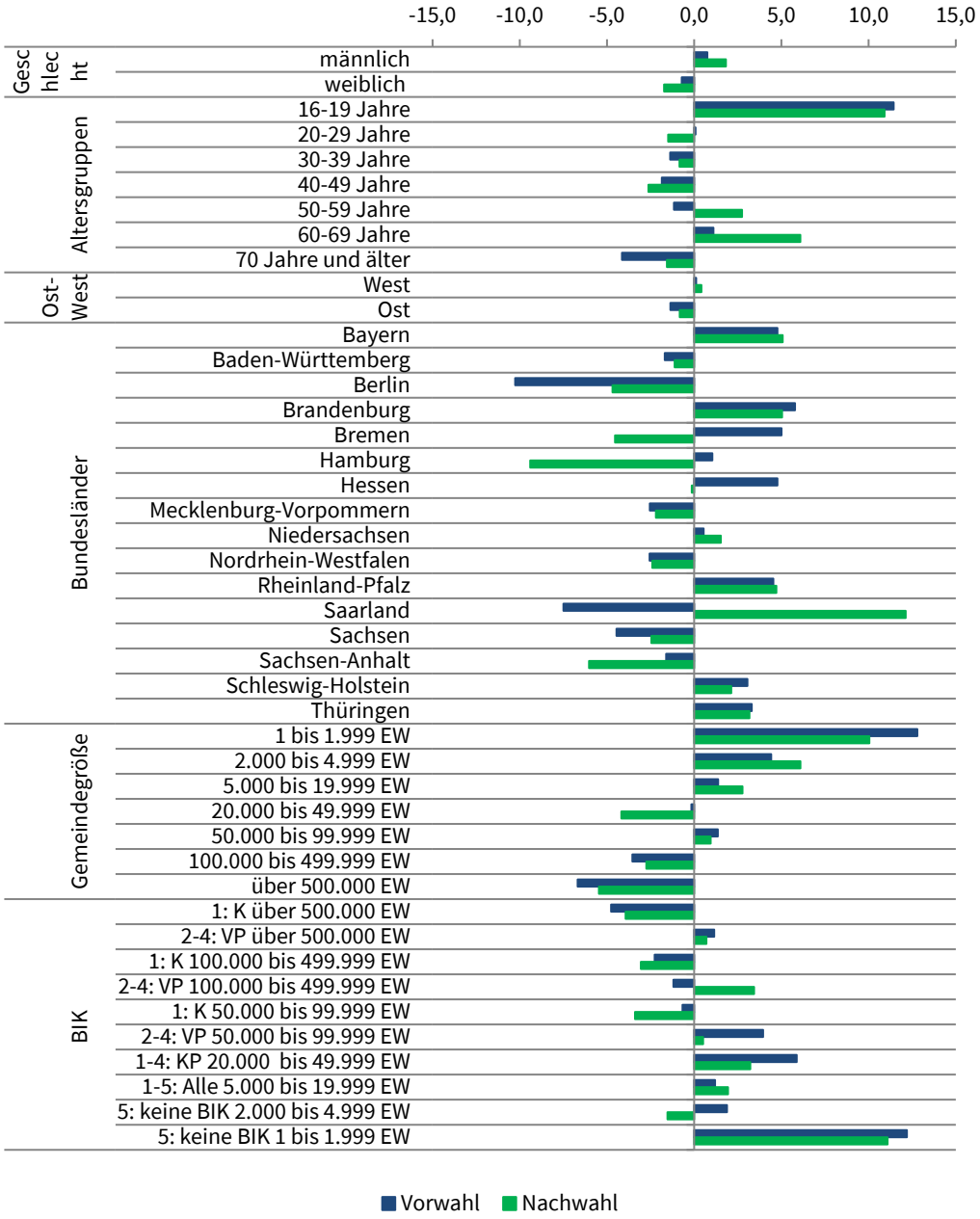


Abbildung 2: Ausschöpfungsquoten (Differenz aus Ausschöpfungsquote verschiedener sozialstruktureller Gruppen minus durchschnittliche Ausschöpfungsquote insgesamt)

Schließlich zeigt sich noch sehr deutlich die Abhängigkeit der Ausschöpfungsquote von der Gemeindegrößeklasse bzw. der BIK-Regionen: Während in Vor- und Nachwahlbefragung in kleinen Gebietskörperschaften bis 19.999 Einwohner/innen die Ausschöpfungsquote überdurchschnittlich hoch ist, fällt diese in Gebieten über 100.000 Einwohner/innen unterdurchschnittlich aus. Für die Qualität der Stichprobe ist dies ein großes Problem, da davon auszugehen ist, dass sich die Meinung, Einstellung und Verhaltensweise von Personen aus ländlichen Gebieten deutlich von denen aus Städten unterscheidet (Stichwort: „rural-urban-divide“, Sixtus, Slupina, Sütterlin, Amberger & Klingholz, 2019). Diese Problematik ist in der Sozialwissenschaft bereits bekannt, weshalb im GLES-Querschnitt bereits 2017 ein „Großstadtzuschlag“ für die Interviewer/innen und z.T. auch die Interviewten bezahlt wurde (Roßteutscher et al., 2019). Die Ergebnisse zeigen jedoch, dass dem Effekt nicht ausreichend entgegengewirkt werden konnte. Es ist vielmehr zu vermuten, dass die Effekte noch stärker gewesen wären, wenn die beschriebene Maßnahme nicht eingesetzt worden wäre.

2.2 Antwortverhalten der Befragten

Nachfolgend verlassen wir die Betrachtungsebene der Bruttodaten und fokussieren das Antwortverhalten der *befragten* Personen. Um Aussagen über die Qualität der Verteilungen machen zu können, vergleichen wir das Antwortverhalten der Befragten mit Datenquellen, bei denen wir davon ausgehen, dass sie den „tatsächlichen“ Wert widerspiegeln und Verteilungen in der Bevölkerung möglichst nahe kommen. Somit kann eine sehr valide und dadurch qualitativ hochwertige Datengrundlage gebildet und als Gradmesser eingesetzt werden.

Zwei Datenquellen kommen hierfür zum Einsatz: (1) die amtliche Statistik (Mikrozensus) und (2) tatsächliches Wahlverhalten bei Bundestagswahlen. Der Mikrozensus ist die größte jährliche Haushaltsbefragung der amtlichen Statistik in Deutschland und unterscheidet sich von anderen Umfragen darin, dass die zufällig ausgewählten Personen zur Auskunft verpflichtet sind, um eine repräsentative und valide Stichprobe der Bevölkerung zu bekommen (Destatis, 2019). Daher eignet sich der Vergleich des Mikrozensus mit den GLES-Querschnittsumfragen sehr gut. Zwar versuchen die GLES-Querschnitte auch möglichst repräsentativ zu sein; da jedoch keine Auskunftspflicht besteht, nehmen – wie in Kapitel 2.1 dargestellt – nicht alle Personen an der Umfrage teil, was wiederum Einfluss auf die Repräsentativität hat. Datengrundlage der nachfolgenden Berechnungen bilden neben den Datensätzen der GLES-Querschnitte (Rattinger, Roßteutscher, Schmitt-Beck, Weßels, Wagner et al., 2019; Rattinger, Roßteutscher, Schmitt-Beck, Weßels, Wolf et al., 2019; Roßteutscher et al., 2019), Auswertungen des Statistischen Bundesamtes für 2017[1] und eigene Berechnungen mit Daten des Mikrozensus für die Jahre 2009 und 2013 (FDZ der statistischen Ämter des Bundes und der Länder, 2009; FDZ der statistischen Ämter des Bundes und der Länder, 2013). Dadurch ist es möglich zu beurteilen, ob und inwiefern sich die Befragungsteilnehmer/innen von der fokussierten Grundgesamtheit stark unterscheiden, was auf Stichprobenqualitätsprobleme hindeuten würde.

Darauf aufbauend wird das tatsächliche Wahlverhalten in Form von Wahlbeteiligung und Stimmabgabe bei Bundestagswahlen mit den Ergebnissen der GLES-Querschnitte 2009, 2013 und 2017 verglichen (Böth & Kobold, 2013; Gisart, 2009; Stemmer, 2017). Schließlich ermöglichen Wahlstudien – anders als viele andere sozialwissenschaftliche Studien – den direkten Vergleich von erhobenen Verhaltensdaten mit dem tatsächlichen Wahlverhalten in Form der Wahlbeteiligung und

[1] Die Auszählungen wurden uns dankenswerterweise vom Statistischen Bundesamt berechnet und per E-Mail zur Verfügung gestellt.

Stimmabgabe. Referenzpunkte des Wahlverhaltens stammen vom Bundeswahlleiter (Bundeswahlleiter, 2018).

Das Ziel des Vergleichs besteht in beiden Fällen darin, die Qualität der GLES-Querschnitte zu beurteilen, indem die GLES-Querschnittsdaten in ihrer möglichst reinen, und somit ungewichteten Form analysiert werden. Deshalb werden ausschließlich notwendige Transformationsgewichte verwendet. Da die Stichprobe der GLES-Querschnitte ein Oversampling der ostdeutschen Bevölkerung hat, muss das Oversampling von Befragten aus Ostdeutschland für alle drei Erhebungsjahre mittels Gewichtung ausgeglichen werden. Die Ziehung der Bruttostichprobe in den GLES-Querschnitten 2009 und 2013 erfolgte auf Haushaltsebene, weshalb die Daten zusätzlich mit einem entsprechenden Transformationsgewicht von einer Haushalts- in eine Personenstichprobe gewichtet werden müssen. Da 2017 eine Registerstichprobe gezogen wurde, ist eine entsprechende Transformation nicht notwendig.

In Analogie zu den GLES-Qualitätsberichten 2009 und 2013 (Blumenberg et al., 2013; Roßmann et al., 2017) werden drei Verfahren angewandt, um Gemeinsamkeiten und Unterschiede der Verteilungen sichtbar zu machen: Hoover-Index, Konfidenzintervalle und Chi-Quadrat-Test.

Hoover-Index

Der Hoover-Index (Hoover, 1936) ermöglicht die Beschreibung von Ungleichverteilungen zwischen zwei Merkmalsverteilungen und gibt den Anteil an, der umverteilt werden müsste, um eine Gleichverteilung zwischen den beiden betrachteten Verteilungen zu erreichen (vgl. Blumenberg et al., 2013, Roßmann et al., 2017). Der Berechnung liegt folgende Formel zugrunde (vgl. Blumenberg et al., 2013: 135):

$$\mathrm{H} = \frac{1}{2}\sum_{i-1}^{N}\left|\frac{E_i}{E_{total}} - \frac{A_i}{A_{total}}\right|.$$

Der Hoover-Index gibt die Abweichung zwischen der Referenzstudie (Mikrozensus/ Bundeswahlleiter) und den GLES-Querschnitten an. E_i/E_{total} gibt dabei den Anteil einer Kategorie in der Referenzstudie an; A_i/A_{total} gib den Anteil aus der GLES-Studie an. Je höher der Hoover-Wert ist, desto größer ist die Umverteilung, die vorgenommen werden muss, um eine Gleichverteilung zu erreichen. Somit sprechen hohe Werte für eine mangelnde Repräsentativität der GLES-Querschnitte.

Vergleich der Mittelwerte inklusive Konfidenzintervallen

Hier wird der „wahre Wert“ – bekannt aus den Mikrozensusdaten oder den Wahlergebnissen – mit den Mittelwerten der GLES-Querschnittsverteilungen inklusive ihren 95-Prozent-Konfidenzintervallen grafisch dargestellt. Als unproblematisch werden Werte angesehen, insofern die „tatsächlichen“ Werte innerhalb der Konfidenzintervalle der GLES-Querschnitte liegen. Problematischer ist es, wenn sich die „tatsächlichen“ Werte außerhalb der Konfidenzintervalle der GLES-Querschnitte befinden.

Chi-Quadrat-Test

Schließlich wird mittels des Chi-Quadrat-Tests festgestellt, ob sich die Verteilungen signifikant voneinander unterscheiden oder nicht. Hierbei werden die beobachteten Werte der GLES-Querschnitte in Bezug zu den erwarteten Häufigkeiten der Referenzstudie gesetzt (vgl. Blumenberg et al., 2013, S. 135).

Nachfolgende Tabelle 2 zeigt die Werte des Hoover-Index und des Chi-Quadrat-Tests für Vorwahl, Nachwahl und Kumulation in den Jahren 2009, 2013 und 2017. Die Darstellung der Mittelwerte und Konfidenzintervalle erfolgt bei der nachfolgenden Darstellung einzelner Merkmale.

Tabelle 2: Hoover-Index und Chi-Quadrat-Werte für verschiedene Merkmale nach Erhebungsjahren differenziert nach GLES-Vorwahlquerschnitt (VW), GLES-Nachwahlquerschnitt (NW) und GLES-Kumulation (Kum.)

		2009			2013			2017		
Merkmal		**VW**	**NW**	**Kum.**	**VW**	**NW**	**Kum.**	**VW**	**NW**	**Kum.**
Geschlecht	Hoover-Index	0,25	1,35	0,54	2,16	0,86	1,53	1,05	3,20	2,11
	Chi2	0,96	0,79	0,91	0,67	0,86	0,76	0,83	0,52	0,67
Alter	Hoover-Index	3,17	3,42	2,83	13,04	14,26	13,63	1,45	1,46	0,86
	Chi2	0,91	0,91	0,94	0,04	0,01	0,03	0,99	0,98	1,00
Bildung	Hoover-Index	4,64	7,30	5,96	8,30	7,10	7,71	12,17	14,58	13,36
	Chi2	0,51	0,24	0,37	0,18	0,29	0,23	0,03	0,01	0,02
Berufliche	Hoover-Index	4,88	7,32	5,72	12,69	10,01	11,33	6,54	4,99	5,76
Stellung	Chi2	0,49	0,22	0,39	0,00	0,02	0,01	0,54	0,51	0,56
Familienstand	Hoover-Index	8,79	8,22	8,51	14,58	12,02	13,14	3,99	3,33	3,61
	Chi2	0,10	0,09	0,11	0,00	0,00	0,00	0,86	0,91	0,91
Haushaltsgröße	Hoover-Index	14,06	11,48	12,66	15,36	14,10	14,74	4,19	5,54	4,85
	Chi2	0,00	0,04	0,01	0,00	0,01	0,00	0,77	0,64	0,72
Wahlbeteiligung	Hoover-Index	2,98	8,59	5,76	8,22	13,17	10,65	14,65	13,71	14,19
	Chi2	0,51	0,06	0,21	0,07	0,00	0,02	0,00	0,00	0,00
Wahlentscheidung	Hoover-Index	7,31	4,43	4,62	6,97	8,10	7,12	7,43	7,42	5,05
	Chi2	0,78	0,87	0,88	0,58	0,54	0,71	0,74	0,70	0,83

2.2.1 Soziodemographische Merkmale

Nachfolgend werden die soziodemographischen Merkmalsausprägungen von Geschlecht, Alter, Bildung, beruflicher Stellung, Familienstand, Haushaltsgröße und Bundesland der GLES-Querschnitte 2009, 2013 und 2017 mit den Daten des Mikrozensus verglichen. Hierzu wurden Abbildungen erstellt, die den anhand des Mikrozensus ermittelten „wahren Wert“ mit den Mittelwerten des GLES-Querschnitts inklusive den 95-Prozent-Konfidenzintervallen darstellen. Wie beschrieben werden Werte als unproblematisch angesehen, insofern die Mikrozensuswerte innerhalb der Konfidenzintervalle der GLES-Querschnitte liegen. Problematisch sind Werte, wenn sich die Mikrozensuswerte außerhalb der Konfidenzintervalle der GLES-Querschnitte befinden.

2.2.1.1 Geschlecht

Zunächst zur Repräsentation von Frauen und Männer in den GLES-Querschnitten: 2009 lag der Anteil der Frauen laut Mikrozensus bei 51,4 Prozent, der der Männer bei 48,6 Prozent. Abbildung 3 zeigt, dass die Konfidenzintervalle der GLES-Querschnitte der Nachwahlbefragung und der GLES-Kumulation 2009 diese Anteilswerte umschließen. In der Vorwahlbefragung liegt der Anteil der Männer allerdings leicht höher als der Referenzwert. 2013 wird der Anteil der Frauen von Vor- und Nachwahlerhebung unter-, der der Männer überrepräsentiert.

Diese Verzerrung weisen auch die GLES-Querschnitte aus 2017 auf, wobei hier die Anteilswerte in der Vorwahlbefragung sehr nahe am „wahren“ Wert des Mikrozensus liegen. Eine Überrepräsentation der Männer und eine Unterrepräsentation der Frauen in der Nachwahlbefragung 2017 sorgen allerdings dafür, dass das Konfidenzintervall der kumulierten Daten die Referenzwerte nicht einschließt. Somit sind Frauen auch im GLES-Querschnitt 2017 unter- und Männer überrepräsentiert.

Diese Abweichungen werden auch von den Hoover-Index-Werten bestätigt: 2009 liegt die Verzerrung für die Nachwahlerhebung und die Kumulation bei unter einem Prozent, während die Abweichungen der Anteile in der Nachwahlbefragung 1,35 Prozentpunkte beträgt. 2013 zeigt sich ein ähnliches Bild für die Vorwahlbefragung (2,16 Prozentpunkte), wobei die Anteile der Nachwahlbefragung nur um 0,86 Prozentpunkte abweichen. Die Nachwahlbefragung 2017 weist mit einem

Wert von 3,20 die deutlichste Abweichung auf, die Anteilswerte der Vorwahl „verfehlen“ die Referenzwerte um 1,05 Prozentpunkte. Die Werte des Chi-Quadrat-Tests sind für keine Befragung signifikant, weshalb die Ungleichverteilungen als weniger problematisch einzustufen sind.

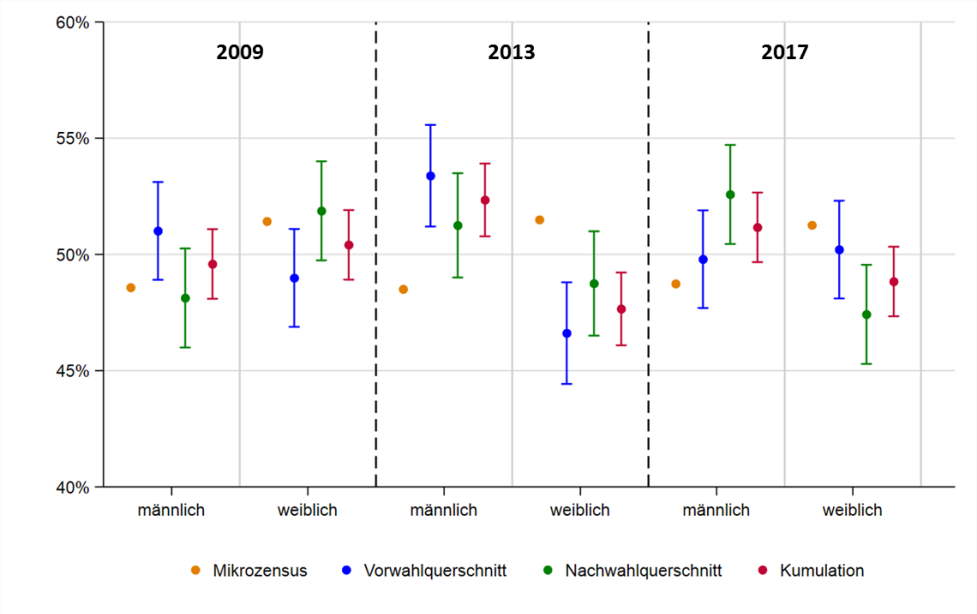


Abbildung 3: Geschlecht

2.2.1.2 Alter

Ein bekanntes Problem bei Face-to-Face-Befragungen ist die Überrepräsentation der älteren Generation (Roßteutscher et al., 2019). Für den GLES-Querschnitt stellt sich die Frage, ob und inwiefern eine Verzerrung hinsichtlich des Alters beobachtet werden kann.

Da die Altersverteilungen des Mikrozensus in den Kategorien „16 bis 29 Jahre“, „30 bis 44 Jahre“, „45 bis 59 Jahre“ sowie „60 Jahre und älter“ vorliegen, wurden entsprechende Altersgruppen mit den Daten der GLES-Querschnitte gebildet und verglichen. Der GLES-Querschnitt 2009 weist in den Anteilswerten der Altersgruppen sehr geringe Abweichungen zu den Referenzwerten auf. Dagegen ist dies 2013 lediglich in der Altersgruppe von „45 bis 59 Jahren“ der Fall. Somit sind die beiden jüngeren Kategorien unter- und die Gruppe der „ab 60-Jährigen“ überrepräsentiert. In den Erhebungen aus 2017 liegt der Referenzwert – mit Ausnahme der „16 bis 29-Jährigen“, die in der Nachwahlerhebung leicht überschätzt werden – im Konfidenzintervall des jeweiligen Anteilswerts.

Der Hoover-Index zeichnet für 2009 und 2017 ein ähnlich positives Bild. Bei diesen Wahlen weichen die Anteilswerte aller Altersgruppen jeweils um ca. ein bis drei Prozentpunkte ab und die Chi-Quadrat-Tests stellen keine signifikanten Unterschiede fest. In der Nachwahlbefragung 2017 ergibt sich sogar „nur“ eine Abweichung von ca. einem Prozentpunkt. 2013 fallen die Hoover-Werte dagegen deutlich höher aus (zwischen 13 und 15 Prozent), was insbesondere der Überrepräsentation der Gruppe der ab „60-Jährigen“ zuzuschreiben ist. Die Unterschiede der Merkmalsverteilung sind beim Chi-Quadrat-Test auch auf dem 0.05-Signifikanzniveau nachzuweisen, was dafürspricht, dass insbesondere die Stichprobenziehung 2013 hinsichtlich der Altersrealisierung als problematisch einzustufen ist.

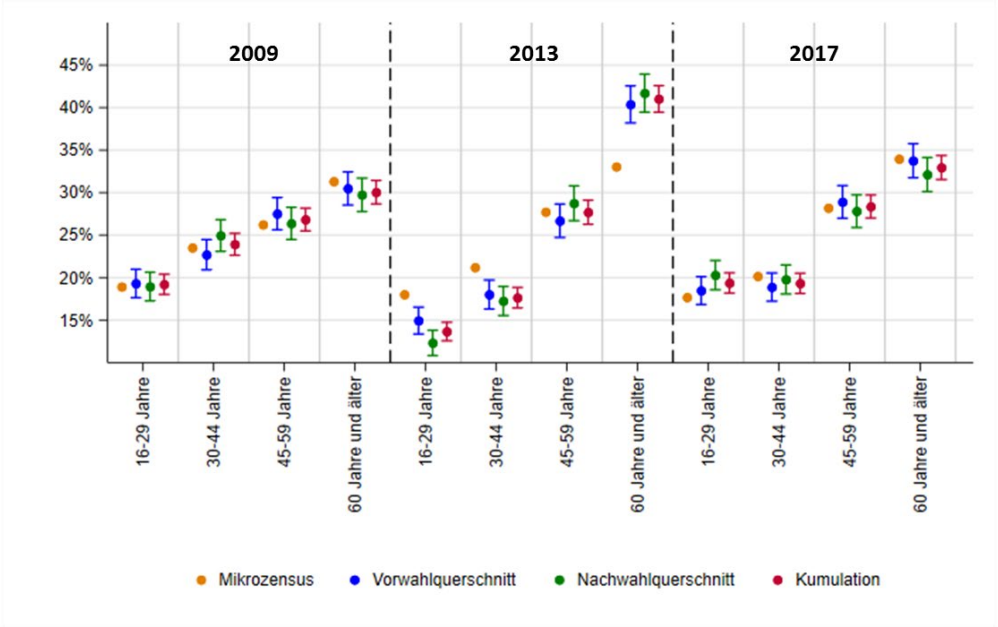


Abbildung 4: Altersgruppen

2.2.1.3 Bildung

Die Antwortkategorien der Frage nach dem höchsten allgemeinbildenden Schulabschluss werden in drei Gruppen aufgeteilt: „niedrige Bildung“ (umfasst „keinen“ oder einen „Hauptschulabschluss“ und äquivalente Abschlüsse sowie die Angabe „bin noch Schüler“), mittlere Reife und äquivalente Abschlüsse werden der „mittlere Bildung“ zugewiesen und Abitur sowie Fachhochschulreife werden zu „hoher Bildung“ zusammengefasst.

Beim Blick über alle drei Erhebungsjahre zeigen sich deutliche Veränderungen in der Zusammensetzung der Bildungsvariablen. 2009 wurde der Anteil von Personen mit niedriger Bildung, bis auf eine leichte Abweichung in der Nachwahl, vergleichsweise gut abgebildet. Personen mit „mittlerer Bildung“ waren dagegen deutlich über- und solche mit hoher Bildung unterrepräsentiert. Vergleichbare Abweichungen zeigen sich auch 2013, wo Befragte mit „mittlerer Bildung“ über- und mit „hoher Bildung“ – letzteres zumindest in der Vorwahlwahlbefragung – unterrepräsentiert sind. Darüber hinaus wurde hier auch der Anteil der Niedriggebildeten unterschätzt.

Im GLES-Querschnitt 2017 verdeutlicht sich ein Problem, mit dem persönliche Interviews aktuell immer mehr zu kämpfen haben. Während die Gruppe mit mittlerem Bildungsstand erstmals adäquat abgebildet werden konnte und der entsprechende Anteilswert des Mikrozensus in den Konfidenzintervallen von Vorwahl- und Nachwahlbefragung und der Kumulation liegt, zeigen sich bei den Hoch- und Niedriggebebildeten deutliche Abweichungen zur tatsächlichen Verteilung des Bildungsstands. Personen mit niedriger Bildung werden stark unter- sowie Personen mit hoher Bildung deutlich überschätzt.

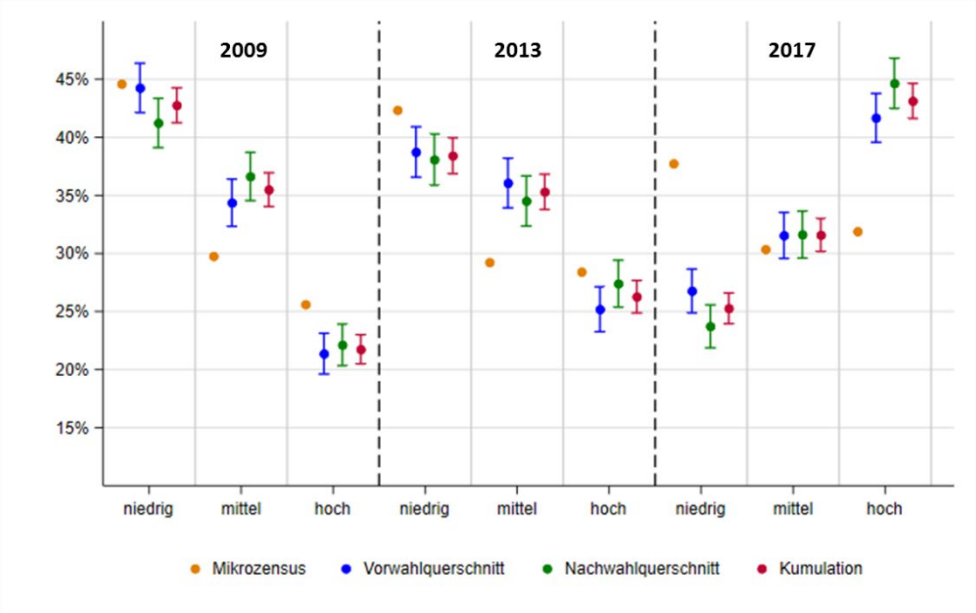


Abbildung 5: Bildung

Diese Abweichungen schlagen sich auch in den jeweiligen Hoover-Werten nieder. Diese liegen in fast allen Studien über 5 Prozentpunkten. Die Nachwahlbefragung 2017 weist hier mit 14,58 den höchsten Abweichungswert auf. Für die GLES-Querschnittskumulation 2017 können sogar Werte in Höhe von 13,36 beobachtet werden. Die Ergebnisse der Chi-Quadrat-Tests weisen die Problematik der Stichprobenrealisierung hinsichtlich Bildung in besonderer Weise 2017 nach. Alle Tests sind auf einem Niveau von 0.05 signifikant. Dies lässt die Schlussfolgerung zu, dass die Bildungsstruktur der Grundgesamtheit in den GLES-Querschnitten 2017 mangelhaft abgebildet wurde und somit Repräsentativitätsprobleme zu diagnostizieren sind.

2.2.1.4 Berufliche Stellung

Neben den zentralen sozialstrukturellen Faktoren Alter, Geschlecht und Bildung werden nachfolgend die berufliche Stellung, der Familienstand und die Haushaltsgröße betrachtet. Bei der beruflichen Stellung liegen die Anteilswerte der GLES-Querschnittsbefragungen in allen Erhebungsjahren tendenziell nahe an den Werten des Mikrozensus. Deutliche Abweichungen zeigt jedoch der Anteil der Arbeiter/innen – 2009 angemessen repräsentiert, 2013 dagegen stark überrepräsentiert, während 2017 in der Vorerhebung leicht überrepräsentiert. Der Anteil der Angestellten wird dagegen in allen Erhebungsjahren adäquat abgebildet.

Die Anteilswerte von Beamten und Selbstständigen in den GLES-Querschnitten von 2013 liegen alle sehr nahe an den entsprechenden Mikrozensuswerten und werden von den jeweiligen Konfidenzintervallen eingeschlossen. 2009 und 2017 zeigt sich eine leichte Überrepräsentation Selbstständiger und auch der Anteil an Beamt/innen wird 2017 leicht überschätzt.

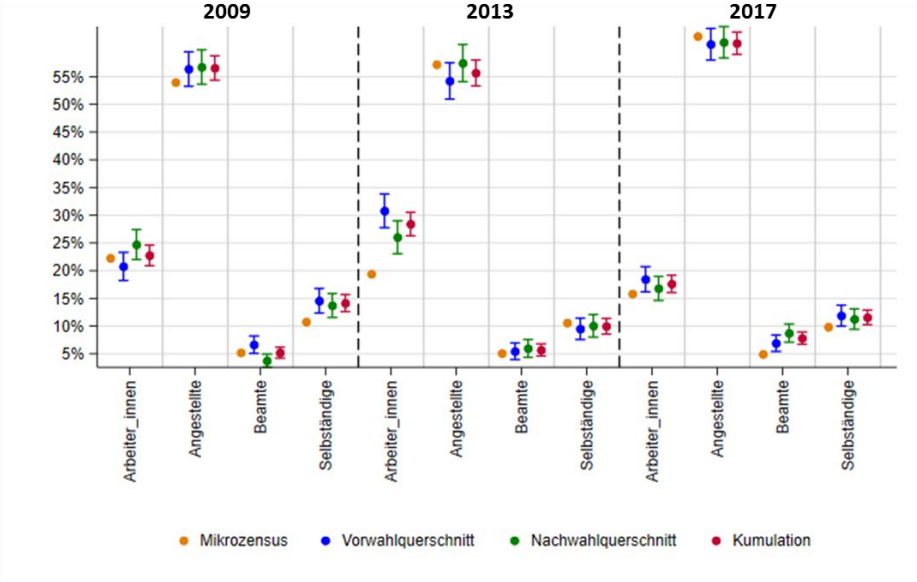


Abbildung 6: Berufliche Stellung

Bei den Hoover-Werten weisen die GLES-Querschnitte 2009 und 2017 Werte um fünf bis sieben Prozentpunkte auf. Deutlich schlechter schneiden die Befragungen 2013 ab. Der Abweichungswert der Vorwahl liegt hier bei 12,69 und der der Nachwahl bei 10,01. Und auch die Chi-Quadrat-Testwerte weisen 2013 auf deutliche Unterschiede in den Verteilungen hin, die die Repräsentativität der Stichprobe diesbezüglich in Frage stellen.

2.2.1.5 Familienstand

Abweichungen zum Mikrozensuswert sind auch in Bezug auf den Familienstand in allen Erhebungsjahren zu beobachten. Hier zeigt sich, dass tendenziell der Anteilswert von verheirateten über- und der von ledigen Personen unterschätzt wird. Es ist anzunehmen, dass hier ein Prädiktor der Teilnahmewahrscheinlichkeit bei persönlichen Interviews deutlich wird, der eng mit dem Familienstand zusammenhängt: die Erreichbarkeit. Eine ledige Person wird aufgrund der höheren Wahrscheinlichkeit, dass sie alleinstehend ist, schlechter zu erreichen sein als der gemeinsame Haushalt eines verheirateten Paars (Groves & Couper, S. 151).

Nichtsdestotrotz lässt sich auch beim Familienstand, verglichen mit den beiden anderen Erhebungsjahren, eine deutliche Verbesserung im GLES-Querschnitt 2017 ausmachen. Die Abweichungen bei den Verheirateten und Ledigen fallen hier weniger stark aus, wobei in der Nachwahlbefragung der Mikrozensus-Anteilswert der ledigen Personen sogar vom Konfidenzintervall eingeschlossen wird.

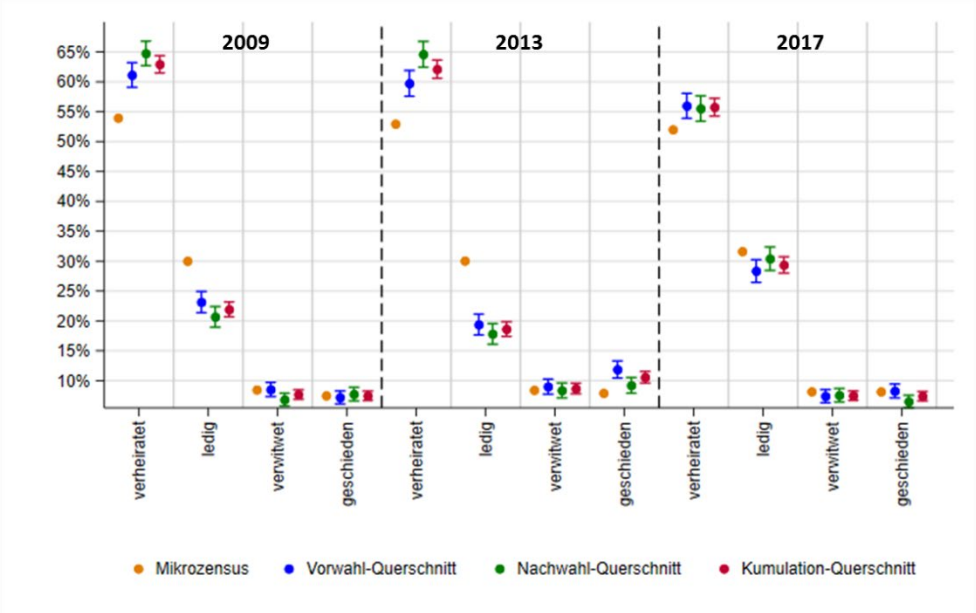


Abbildung 7: Familienstand

Weniger deutlich fallen die Abweichungen bei den Verwitweten bzw. Geschiedenen aus. Hier treffen die GLES-Querschnitte 2009 und 2017 den tatsächlichen Anteilswert relativ gut, während 2013 nur der Anteil der Geschiedenen in der Vorwahlbefragung (und dadurch auch in der Kumulation) überschätzt wird.

Diese Ergebnisse spiegeln sich auch in den Hoover- und Chi-Quadrat-Werten wider. Während die Verteilungen in Vor- und Nachwahlerhebung 2009 und 2017 nicht signifikant von den Mikrozensusverteilungen abweichen, schneidet der GLES-Querschnitt 2013 mit signifikanten Unterschieden zum Mikrozensus schlechter ab. Die Verbesserung im GLES-Querschnitt 2017 gegenüber den vorherigen Befragungen zeigt sich auch in den Hoover-Werten. Während beide Erhebungen 2013 mit 14,58 bzw. 12,02 relativ hohe Werte ausweisen, liegen die Werte 2009 mit 8,79 bzw. 8,22 deutlich darunter. 2017 weist mit 3,99 bzw. 3,33 nochmals eine deutliche Verbesserung der Hoover-Werte auf.

2.2.1.6 Haushaltsgröße

Ein ähnliches Bild wie beim Familienstand zeigt sich auch bei der Haushaltsgröße. In den GLES-Querschnitten 2009 und 2013 werden Zweipersonenhaushalte über- und Einpersonenhaushalte unterschätzt. Lediglich die Konfidenzintervalle der beiden Vorwahlbefragungen schließen den wahren Wert von Haushalten mit nur einer Person ein. Deutlicher fällt diese Unterschätzung 2017 aus, wo sich der Abstand zum Mikrozensuswert noch größer zeigt. Hier wird jedoch der Anteil von Zweipersonenhaushalten in beiden Erhebungen adäquat abgebildet. Es ist anzunehmen, dass diese Verzerrung mit der Erreichbarkeit von Personen zusammenhängen könnte, da Personen in Einpersonenhaushalten schwerer zu erreichen sind als Personen in Haushalten mit mehreren Haushaltsmitgliedern (Brick & Williams, 2013, S. 49).

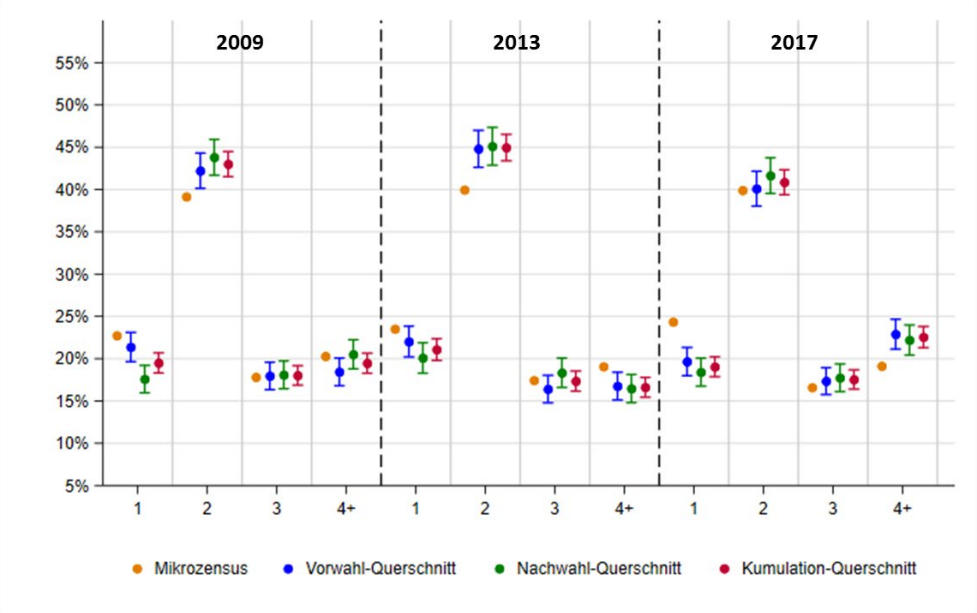


Abbildung 8: Haushaltsgröße

2009 werden die Anteile von Haushalten mit drei bzw. mindestens vier Personen relativ gut getroffen. Die anderen beiden Befragungen bilden zwar Dreipersonenaushalte gut ab, zeigen dagegen bei den Haushalten mit vier oder mehr Personen teils größere Abweichungen. Auffällig ist hierbei, dass der Anteil 2013 unter- und 2017 überschätzt wird.

Mit Blick auf die Hoover- und Chi-Quadrat-Werte schneidet auch hier der GLES-Querschnitt 2017 insgesamt deutlich besser ab. Die größten Hoover-Werte weist 2013 mit 15,36 und 14,10 auf. Auch die Erhebungen 2009 liegen mit 14,06 und 11,48 deutlich über denen von 2017 (4,19 und 5,54). Dieses Bild zeigt sich auch bei den Chi-Quadrat-Tests, wo 2009 und 2013 mit teils höchstsignifikanten Werten von den Mikrozensus-Verteilungen abweichen, während bei den Erhebungen 2017 insgesamt keine Unterschiede zur Referenzverteilung angenommen werden können.

2.2.1.7 Exkurs: Merkmalskombinationen

Bei der Repräsentation von Stichproben ist es nicht nur wichtig, dass zentrale Merkmale in ihrer univariaten Verteilung ähnlich dem „wahren“ Wert sind, sondern auch, dass die Merkmale in ihrer Kombination der Repräsentation in der Grundgesamtheit entsprechen. Einfach ausgedrückt: In der Stichprobe sollen Frauen aller Altersgruppen und Bildungsniveaus enthalten sein und nicht vermehrt nur niedrig Gebildete, was eine Verzerrung der Stichprobe bedeuten würde. Aus diesem Grund werden nachfolgend die drei zentralen Merkmalsverteilungen Alter, Geschlecht und Bildung kombiniert betrachtet.

Es wurden für jedes Jahr die Differenzen der Anteile zwischen der jeweiligen Ausprägung in der GLES-Querschnittskumulation minus den Werten im Mikrozensus berechnet. Werte im positiven Bereich bedeuten, dass die jeweilige Gruppe im GLES-Querschnitt überrepräsentiert ist; Werte im negativen Bereich bedeuten eine Unterrepräsentation der jeweiligen Gruppe.

Nacheinander erfolgt nun die Betrachtung der Kombinationen Geschlecht und Alter, Geschlecht und Bildung und Alter und Bildung. Zunächst zur Kombination Geschlecht und Alter: Für den GLES-Querschnitt 2017 zeigt sich eine sehr gute Repräsentation. Problematischer ist die Repräsentation 2009 und besonders problematisch 2013. Sowohl 2009 als auch 2013 sind ältere Männer überre-

präsentiert – 2013 sogar um satte 9,4 Prozentpunkte. Frauen waren 2013 v.a. in den Altersgruppen der 16 bis 29-jährigen unterrepräsentiert und 2009 waren v.a. die über 60-jährigen Frauen in geringeren Maßen bereit, an einem Interview teilzunehmen. Der Überrepräsentation von über 60-jährigen Männern steht eine Unterrepräsentation in den jüngeren Altersgruppen (v.a. 30 bis 59-Jährigen) in den Jahren 2009 und 2013 gegenüber. Insbesondere die immense Überrepräsentation von Männern 2013 muss aus Stichprobensicht als besonders problematisch eingestuft werden.

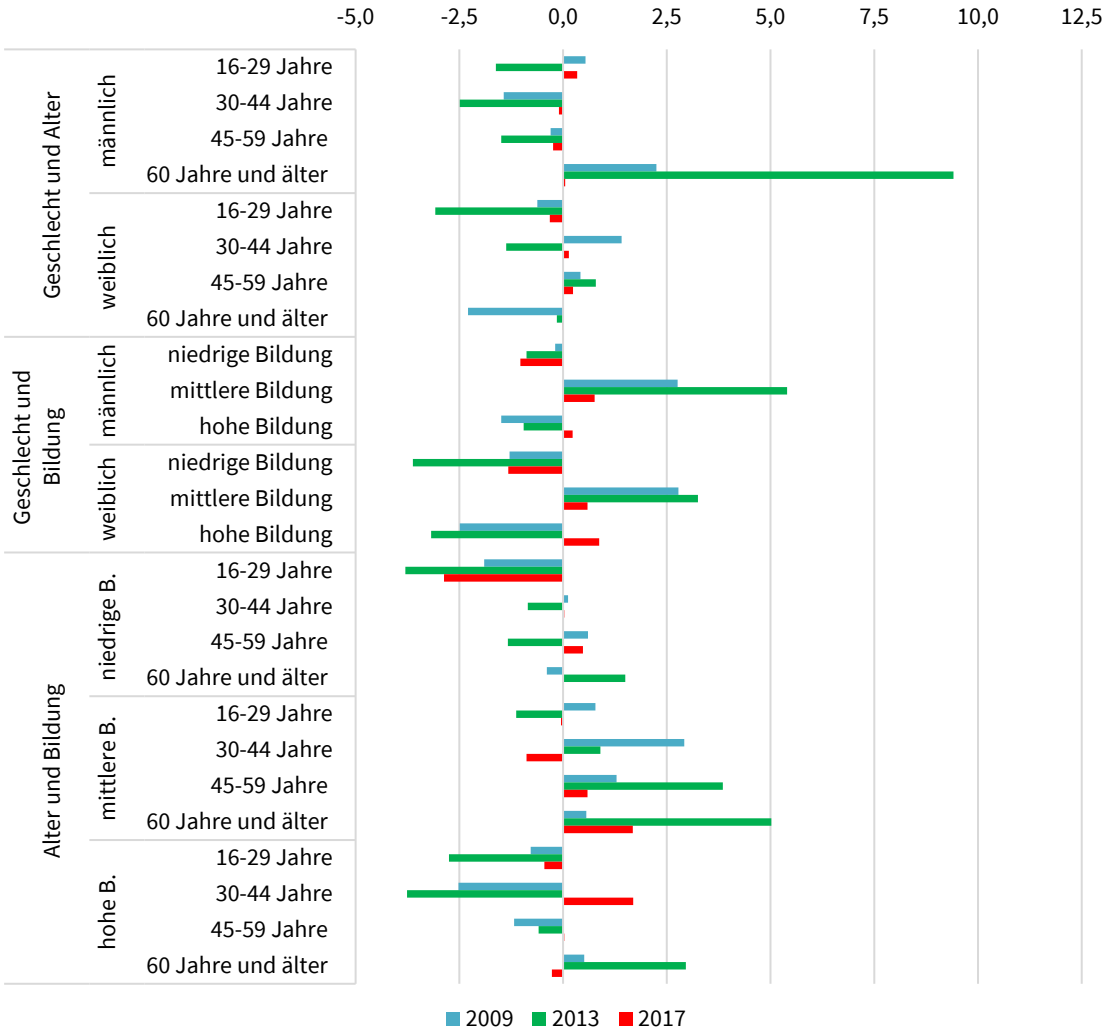


Abbildung 9: Merkmalskombinationen

Über alle Befragungszeitpunkte hinweg können Schwierigkeiten in der Stichprobenrealisierung bezüglich der Bildung beobachtet werden. In Bezug auf Bildung und Geschlecht zeigt sich, dass sowohl Frauen als auch Männer mit niedriger Bildung unterrepräsentiert waren, wobei dies bei den Frauen in stärkerem Maße zu beobachten ist. Personen mit mittlerer Bildung sind in allen drei Jahren bei Männern wie Frauen überrepräsentiert. Jedoch zeigt sich auch hier, dass 2013 besonders viele Männer mit mittlerer Bildung verstärkt an den Interviews teilgenommen haben. Auch bei der hohen Bildung sind weniger starke kombinierte Effekte als Bildungseffekte beider Geschlechter zu beobachten. Während Personen mit hoher Bildung 2017 leicht überrepräsentiert waren,

waren diese 2009 und 2013 unterrepräsentiert. Auch hier zeigt sich, dass Frauen mit hoher Bildung in besonderen Maßen unterrepräsentiert sind.

Schließlich noch zur Kombination von Alter und Bildung. Bei der Betrachtung der niedrig Gebildeten fällt auf, dass die Gruppe, der unter 29-Jährigen bei allen drei Wahlen unterrepräsentiert ist: am stärksten 2013 gefolgt von 2017 und am wenigsten stark 2009. In den Altersgruppen über 30 Jahren fallen die Unterschiede deutlich geringer aus. Wie oben bereits gezeigt wurde, sind die Personen mit mittlerer Bildung tendenziell überrepräsentiert. Doch auch hier kann keine Gleichverteilung über die Altersgruppen beobachtet werden: 2009 waren in besonders starkem Maße die 30 bis 44-Jährigen überrepräsentiert; 2013 waren es die ab 45-Jährigen. Für 2017 ist zu beobachten, dass die Personen mit mittlerer Bildung unter 44 Jahre tendenziell unter- und die über 45 Jahren überrepräsentiert sind. Bei den hochgebildeten Teilnehmern/innen ist zu erkennen, dass diese Gruppe 2009 und 2013 bis 59 Jahren unterrepräsentiert war, während die hochgebildeten über 60 Jahre überrepräsentiert waren. 2017 zeigt sich, dass die Personen mit höherer Bildung überproportional häufig der Altersgruppe der 30 bis 44-Jährigen angehören.

2.2.2 Wahlverhalten

Im Vergleich zu vielen anderen Umfragen hat die empirische Wahlforschung einen (vermeintlichen) Vorteil, da der Bundeswahlleiter kurz nach den Wahlen die exakten Wahlergebnisse hinsichtlich Wahlbeteiligung und Parteiwahl veröffentlicht (2009: Gisart, 2009; 2013: Böth & Kobold, 2013; 2017: Stemmer, 2017). Dadurch existieren – wie bereits dargestellt – zwei Gradmesser, um die Qualität der Stichprobe zu bewerten:

- Vergleich des tatsächlichen Wahlergebnisses mit der Wahlentscheidungsfrage in der Umfrage und
- Vergleich der tatsächlichen Wahlbeteiligung mit der in der Umfrage angegebenen Wahlbeteiligung.

Die Veröffentlichung der Wahlergebnisse ist nicht nur für kommerzielle Umfrageinstitute (wie z.B. Infratest oder die Forschungsgruppe Wahlen) ein wichtiger Maßstab, um die Qualität ihrer Arbeit zu messen. Auch die empirisch-wissenschaftliche Wahlforschung muss sich mit der Frage konfrontieren, inwiefern die erhobenen Daten das tatsächliche Wahlverhalten widerspiegeln.

2.2.2.1 Wahlteilnahme

Ein Problem zahlreicher Wahlumfragen ist, dass der Anteil der Nichtwähler/innen meist niedriger ausfällt als der tatsächliche Anteil bei Wahlen (Voogt & Saris, 2005; Sciarini & Goldberg, 2015). Zwei Ursachen liegen diesem Phänomen zugrunde: Erstens ist es möglich, dass Nichtwähler/innen an den Befragungen seltener teilnehmen. Erklärt werden kann das durch das geringere Interesse an politischen Themen, was in einer Verweigerung der Interviewteilnahme mündet. Zweitens kann sozial erwünschtes Verhalten auftreten. Da die Wahlteilnahme gesellschaftlich als erwünscht gilt und eine befragte Person sich im persönlich-mündlichen Interview entsprechend verhalten möchte, gibt er im Interview an, an der Wahl teilzunehmen, obwohl er es nicht vorhat zu tun bzw. nicht getan hat (Hugi, 2014; Sciarini & Goldberg, 2015).

Die Abbildung 10 zeigt in Analogie zu den oben gezeigten Darstellungen die Wahlbeteiligung im GLES-Vorwahl- und Nachwahlquerschnitt sowie in der Kumulation in den Jahren 2009, 2013 und 2017. Es ist gut zu erkennen, dass auch in der GLES zu allen drei Befragungszeitpunkten eine Überschätzung der Wahlbeteiligung zu beobachten ist. Zusätzlich ist zu erkennen, dass sich die Überschätzung von Wahljahr zu Wahljahr erhöht. Zwar ist die tatsächliche Wahlbeteiligung 2017 ange-

stiegen und belief sich auf 76,2 Prozent. In den GLES-Querschnitten wurden sie dennoch um über 12 Prozentpunkte überschätzt.

Auch die Werte des Hoover-Indexes zeigen die (zunehmende) Ungleichheiten in der Verteilung deutlich: 2009 beläuft sich der Hoover-Index in der Kumulation auf 5,78, steigt 2013 auf 10,65 und lag 2017 bei 14,19. Schließlich weisen auch der hochsignifikante Chi-Quadrat-Test 2017 und der signifikante Test 2013 auf die wachsenden Unterschiede der Verteilungen hin.

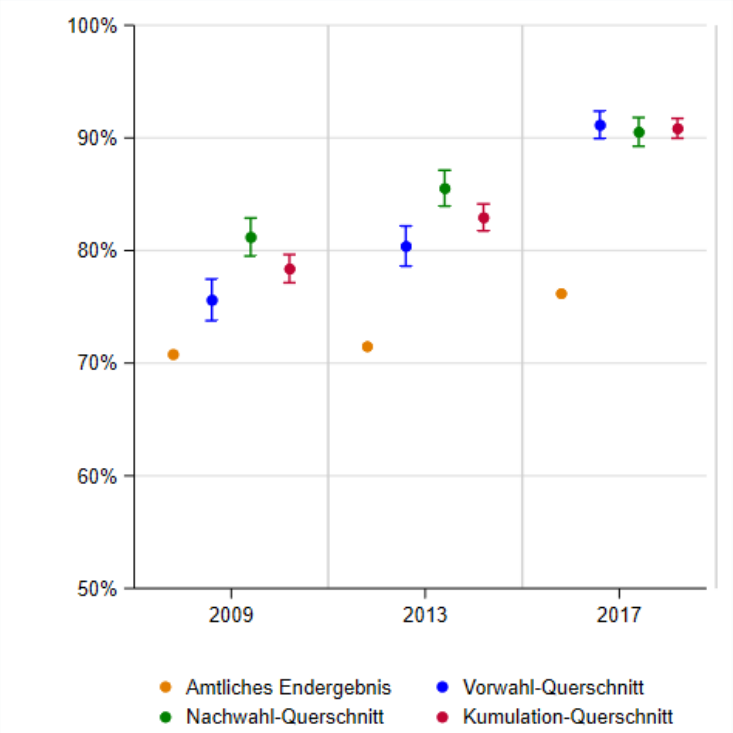


Abbildung 10: Wahlbeteiligung

Diese Problematik konfrontiert die gesamte Gruppe der Wahlforscher/innen, die sich mit der Frage der Wahlbeteiligung befasst, nicht nur mit Repräsentativitäts- sondern auch mit Fallzahlproblemen. Eine Aufgabe der zukünftigen Wahlforschung wird es daher sein, dieses größer werdende Problem theoretisch und umfragemethodisch systematisch zu analysieren, um darauf aufbauend Lösungsstrategien zu entwickeln, die die Verzerrung reduzieren.

2.2.2.1 Wahlentscheidung

Für die Wahlforschung besonders relevant ist schlussendlich die Frage, wie gut die Umfragen die tatsächliche Wahlentscheidung erfasst haben. Gibt es hier Unterschiede zum tatsächlichen Wert und wenn ja, sind erstens Unterschiede zwischen Vor- und Nachwahl zu erkennen und zweitens, welche parteispezifischen Unterschiede können beobachtet werden? Die nachfolgende Abbildung 11 zeigt die Unterschiede im Zweitstimmenanteil von den jeweiligen GLES-Umfragen zu den tatsächlichen Ergebnissen.

Zunächst zu den größeren Parteien CDU/CSU und SPD: Die CDU/CSU-Wahl konnte 2009 und 2013 vom GLES-Querschnitt gut abgebildet werden. 2017 zeigen sich jedoch Abweichungen der GLES-Querschnitte von dem tatsächlichen Wahlverhalten: Die Vorwahl überschätzt und die Nachwahl unterschätzt tendenziell die Wahlentscheidung zugunsten der CDU/CSU. Auch bei der Wahl 2009

und 2013 können Unterschiede bei der SPD dahingehend identifiziert werden, dass das Wahlverhalten tendenziell zugunsten der SPD überschätzt wird.

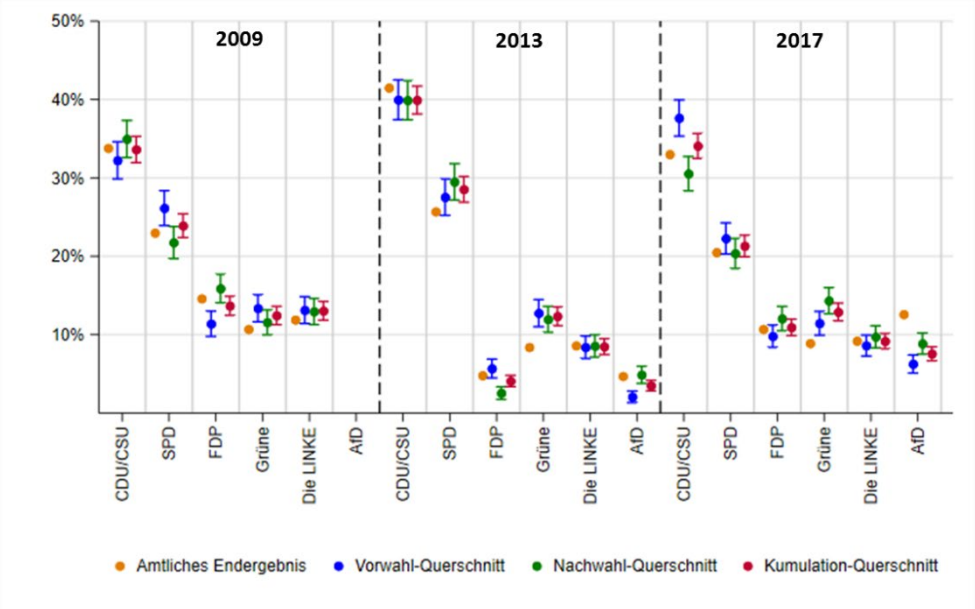


Abbildung 11: Zweitstimmenanteile

Abweichungen sind auch bei den kleineren Parteien zu beobachten. Es zeigt sich, dass die FDP in der Vorwahl- und der Nachwahlbefragung jeweils entgegengesetzt unter- und überschätzt wird, was bei der Kumulation jedoch nicht mehr ins Gewicht fällt. Die Grünen werden tendenziell eher überschätzt. D.h. in der realisierten Stichprobe sind Wähler/innen der Grünen Partei tendenziell überrepräsentiert. Nicht in dem starken Ausmaß, jedoch in der Tendenz ist insbesondere 2009 eine Überrepräsentation auch von Linken-Wähler/innen zu beobachten. Schließlich wurde die AfD in der Vorwahl 2013 unterschätzt, was jedoch nicht verwunderlich ist, da die AfD erst im Februar 2013 gegründet wurde und sich bei einigen Wähler/innen erst kurz vor der Bundestagswahl 2013 zu einer wirklichen Alternativen entwickelt hat. Eine mögliche Erklärung kann daher sein, dass zu Beginn der Datenerhebung acht Wochen vor der Bundestagswahl 2013 für einige Befragte die AfD keine Alternative darstellte, sich dies bis zur Bundestagswahl durch zunehmende Präsenz geändert hat. Mit dem Wandel der AfD von einer eurokritischen Partei zu einer rechten Alternative hat sich auch die Repräsentation der Wählergruppe erhöht (Bieber, Roßteutscher & Scherer, 2018). Dennoch waren auch 2017 AfD-Wähler/innen sowohl in der Vor- als auch in der Nachwahlbefragung stark unterrepräsentiert. Inwiefern dies sozialer Erwünschtheit, Unterrepräsentation von AfD-Wähler/innen oder Interviewereffekten in der realisierten Stichprobe geschuldet ist, kann an dieser Stelle nicht geklärt werden.

In Bezug auf die beiden anderen Kriterien – Hoover-Index und Chi-Quadrat-Test – scheinen die Merkmalsunterschiede weniger problematisch zu sein, da die Unterschiede weder signifikante noch hohe Werte aufweisen. Dennoch muss an dieser Stelle betont werden, dass gerade im Falle der kleinen Parteien nicht ganz exakte Merkmalsverteilungen für die nachfolgende Analyse immense Probleme bedeuten können, die sowohl in der Fallzahlproblematik als auch in der unzureichenden Repräsentativität der Parteiwähler/innen liegen kann. Daher sollten Wahlstudien sich das Ziel setzen sowohl hinsichtlich der Wahlbeteiligung als auch des Wahlverhaltens möglichst exaktes Verhalten zu erfassen, um der wissenschaftlichen Community Daten für angemessene Analysen anbieten zu können.

3 Feldverlauf

Ein homogener Feldverlauf ist für die Datenqualität von Wahlstudien sehr bedeutend. Im Vergleich zu zahlreichen anderen persönlich-mündlichen Interviews sind die Feldzeiten der GLES-Querschnitte sehr kurz: Die Vorwahlerhebung findet in den acht Wochen vor der Bundestagswahl statt und endet am Tag vor der Wahl. Eine Verlängerung der Feldzeit ist nicht möglich, da nicht zuletzt das Frageprogramm der Situation vor der Wahl angepasst ist. Ähnlich verhält es sich in der Nachwahlerhebung: Der Feldstart fällt auf den ersten Tag nach der Wahl und es wird ebenso eine Feldzeit von acht Wochen angestrebt, wobei es grundsätzlich möglich ist, diese in geringfügigem Maße zu verlängern. Jedoch ist auch dies in Wahlstudien problematisch, da das zu beobachtende Verhalten (Wahlentscheidung) dann sehr weit vom Untersuchungszeitpunkt entfernt liegt und dadurch die Erinnerung der Bürger/innen erstens Lücken aufweisen oder zweitens eine Anpassung des (vermeintlichen) Verhaltens an die aktuellen politischen Entwicklungen stattfinden kann. Daher ist eine engmaschige Feldkontrolle sehr wichtig, um möglichst frühzeitig Maßnahmen in die Wege leiten zu können, um Probleme in der Feldarbeit zu korrigieren und die Interviews im vorgegebenen Zeitplan zu realisieren.

3.1 Allgemeiner Überblick

Die zu realisierenden Fallzahlen werden in der GLES vorgegeben. Es sollen in den acht Wochen vor und in den acht Wochen nach der Bundestagswahl jeweils 2100 Interviews realisiert werden. Grundsätzlich strebt die wissenschaftliche Community eine Normalverteilung der realisierten Interviews über die gesamte Vor- und Nachwahlfeldphase an. Das bedeutet, dass in zeitlicher Nähe zum tatsächlichen Wahltag möglichst viele Interviews realisiert werden und zu Beginn der Feldzeit der Vorwahl und am Ende der Feldzeit der Nachwahl weniger.

Die nachfolgende Abbildung 12 zeigt die realisierten Interviews 2009, 2013 und 2017 in Form von Balkendiagrammen. Es ist deutlich zu erkennen, dass eine Normalverteilung über die Feldphase des GLES-Vor- und Nachwahlquerschnitts weder 2009, noch 2013, noch 2017 erreicht werden konnte. Vielmehr können Annäherungen an Normalverteilungen jeweils separat für die Vor- und die Nachwahlerhebungen beobachtet werden. Besonders auffällig ist die niedrige Anzahl an durchgeführten Interviews 2009. In den zwei Wochen vor und in den zwei Wochen nach der Wahl – die Wochen, in denen am meisten Interviews hätten durchgeführt werden sollen – wurden vergleichsweise wenig Interviews realisiert.

Einen deutlich besseren Verlauf nimmt die Vorwahlbefragung 2013, in welcher die Anzahl der Interviews in der Woche vor der Wahl besonders hoch ausfällt. Dieses Hoch wird jedoch von einem starken Tief begleitet. In der Woche nach der Bundestagswahl 2013 wurden nur zwei Interviews durchgeführt und auch in der darauffolgenden Woche wurden nur 56 realisiert. Erst in der sechsten, siebten und achten Woche nach der Bundestagswahl 2013 konnten durchschnittliche Interviewzahlen von ca. 270 Interviews pro Woche durchgeführt werden. Da der Verlauf in der Nachwahl 2013 so schleppend verlief, wurde die Feldzeit von acht Wochen auf insgesamt 14 Wochen verlängert – nicht zuletzt um die vertraglich vereinbarten Interviews zu realisieren und der wissenschaftlichen Community entsprechend viele Fälle zur Verfügung zu stellen – auch wenn der Erhebungszeitpunkt dieser Fälle vom Wahlereignis weit entfernt liegt und davon auszugehen ist, dass erhebliche Qualitätseinbußen damit einhergehen. Somit ist aus Qualitätsaspekten der Feldverlauf der Nachwahlbefragung von 2013 als besonders problematisch einzustufen.

Ein deutlich homogenerer Feldverlauf – wenn auch nicht einer Normalverteilung folgend – kann anlässlich der Bundestagswahl 2017 beobachtet werden: Nachdem in der ersten Woche nur 131 Interviews realisiert werden konnten, stieg die Anzahl der realisierten Interviews in der zweiten, dritten und vierten Woche deutlich an und konnte das Niveau von mindestens 240 Interviews pro Woche in der Vorwahl durchgehend halten. Dennoch kann auch 2017 die geringen Fallzahlen in der ersten Woche nach der Bundestagswahl beobachtet werden, wobei sich die Werte dann ab der zweiten Woche erholten, jedoch in Woche sechs und sieben erneut deutlich absanken, weshalb auch hier die Feldzeit um zwei Wochen verlängert werden musste. Somit kann auch der Feldverlauf der Nachwahlbefragung 2017 als partiell problematisch eingestuft werden, wenn auch nicht so stark wie der Verlauf 2013.

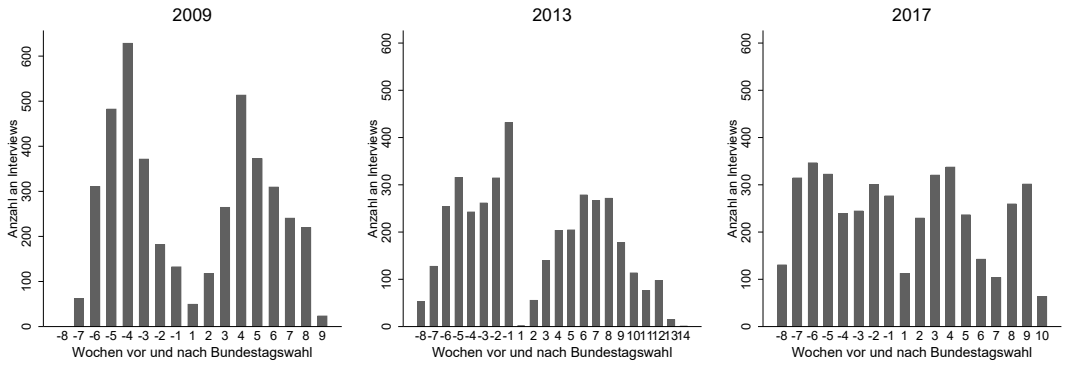
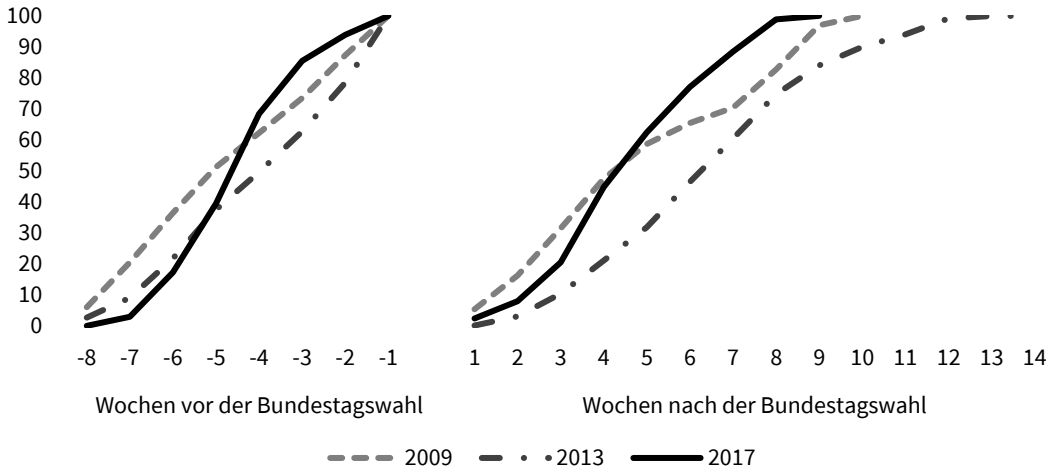


Abbildung 12: Feldverlauf 2009, 2013 und 2017

Die Trägheit der Fallentwicklung in den Nachwahlerhebungen 2013 und 2017 ist auch der nachfolgenden Abbildung 13 zu entnehmen. Diese stellt den kumulierten, prozentualen Anteil an realisierten Interviews pro Woche dar. Während die Vorwahlstudien 2009, 2013 und 2017 tendenziell ähnliche Verläufe zeigen, ist bei den Nachwahlstudien der bereits berichtete schleppende Verlauf 2013 zu beobachten. Ebenso ist für 2017 ein Plateau von Woche fünf bis sieben zu erkennen.

Für das zukünftige Fieldwork-Monitoring sollten daher in Bezug auf den allgemeinen Feldverlauf zwei Dinge festgehalten werden: Erstens, konnte eine Normalverteilung der Interviews von Beginn der Vorwahlerhebung bis Ende der Nachwahlerhebung in keiner der drei GLES-Querschnittsstudien beobachtet werden, was möglicherweise auch mit der getrennten Stichprobenziehung und der schwereren Erreichbarkeit der Interviews am Beginn und am Ende der Feldphase zusammenhängt. Zweitens ist ein schleppender bzw. einschlafender Feldverlauf in zwei der drei Nachwahlbefragungen zu beobachten, die darauf hindeuten, dass mit mangelndem Druck – kein hartes Feldende aufgrund des Wahltags – das Leistungsniveau der Institute sinkt und es daher im Fieldwork-Monitoring der Nachwahl besonders wichtig ist, diese Trägheit einzuplanen und frühzeitig Gegenmaßnahmen zu entwickeln, die einen unverzüglichen Feldstart realisieren.



Anmerkung: nach Blumenberg & Adewuyi, 2017:35, erweitert um Bundestagswahl 2017.

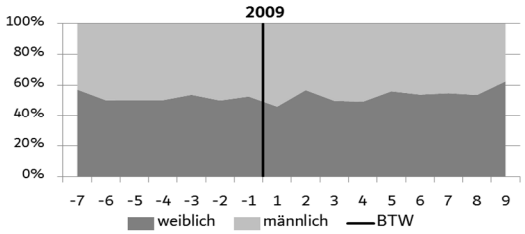
Abbildung 13: Feldverlauf der GLES-Vor- und Nachwahlquerschnitte im Vergleich (in %)

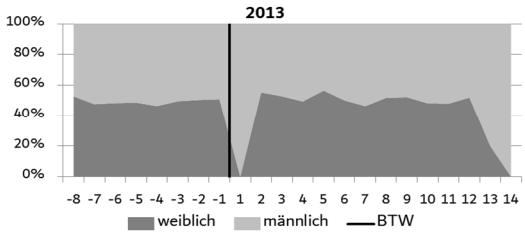
3.2 Sozialstrukturelle Unterschiede

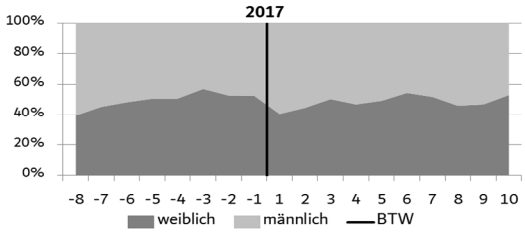
Zusätzlich zu den realisierten Interviews werden die Verteilungen ausgewählter sozialstruktureller Merkmale über den jeweiligen Erhebungszeitraum als Gütekriterium des Feldverlaufs herangezogen, wobei eine möglichst gleiche Verteilung der betrachteten Merkmale über die Feldzeit hinweg ein positives Qualitätsmaß darstellt. Zwar wurde eine solche Gleichverteilung während der Erhebung nicht vertraglich gefordert oder mittels Qualitätskontrollen überprüft – dennoch könnten starke Schwankungen dieser Verteilung aufgrund von Wahlkampfeffekten oder medialer Berichterstattung zu Stichprobenverzerrungen führen. Dies könnte beispielsweise auftreten, wenn ältere Personen vornehmlich am Beginn der Befragung interviewt werden und es kurz vor der Wahl zu starken Diskussionen von Rentenreformen kommt, die massive Auswirkungen auf insbesondere die älteren Gruppierungen hätte, die bei der Befragung jedoch nicht mehr in dem Ausmaß berücksichtigt werden, da die Interviews mit den älteren Personen weitgehend abgeschlossen sind. Daher werden im Folgenden die bereits in 2.2.1 untersuchten soziodemografischen Merkmale Geschlecht, Alter, Bildung, berufliche Stellung, Familienstand und Haushaltsgröße hinsichtlich ihrer Verteilung über den gesamten Feldverlauf mittels Flächendiagrammen betrachtet.

In der Gesamtschau sind – mit einer Ausnahme – tendenziell paralell laufende Flächen zu beobachten, was auf eine homogene Feldbearbeitung hindeutet. 2013 zeigt der Feldverlauf jedoch bei allen Merkmalen in der Woche nach dem Wahltag und am Ende der Nachwahlfeldzeit deutliche Einbrüche. Dies ist jedoch auf die sehr geringe Zahl an realisierten Interviews in den entsprechenden Wochen zurückzuführen. Wie in 3.1 berichtet, wurden in der Woche nach der Wahl (Kalenderwoche 39) nur zwei Interviews und in der 14. Woche (52. Kalenderwoche) lediglich ein Interview mit jeweils einem männlichen Befragten geführt, was die deutlichen Ausschläge in diesen Wochen bei den betrachteten Merkmalen erklärt.

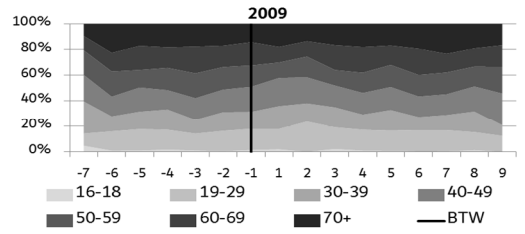
Geschlecht

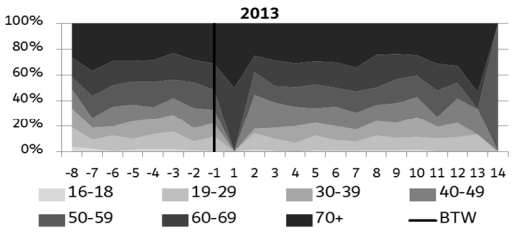


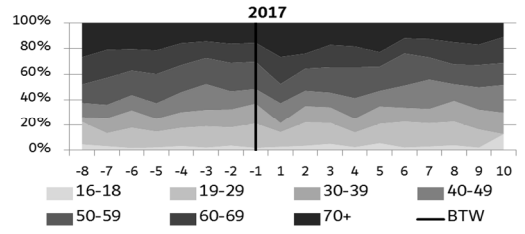




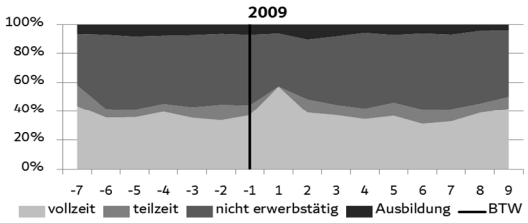
Alter

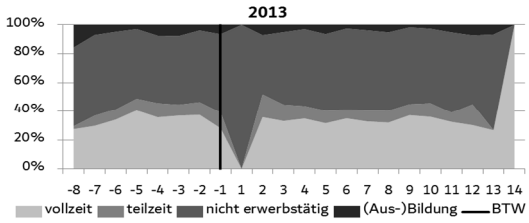


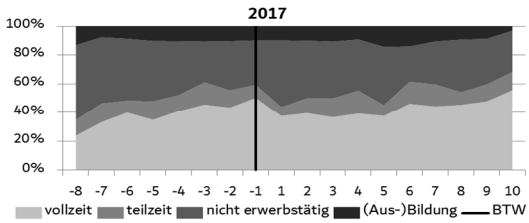




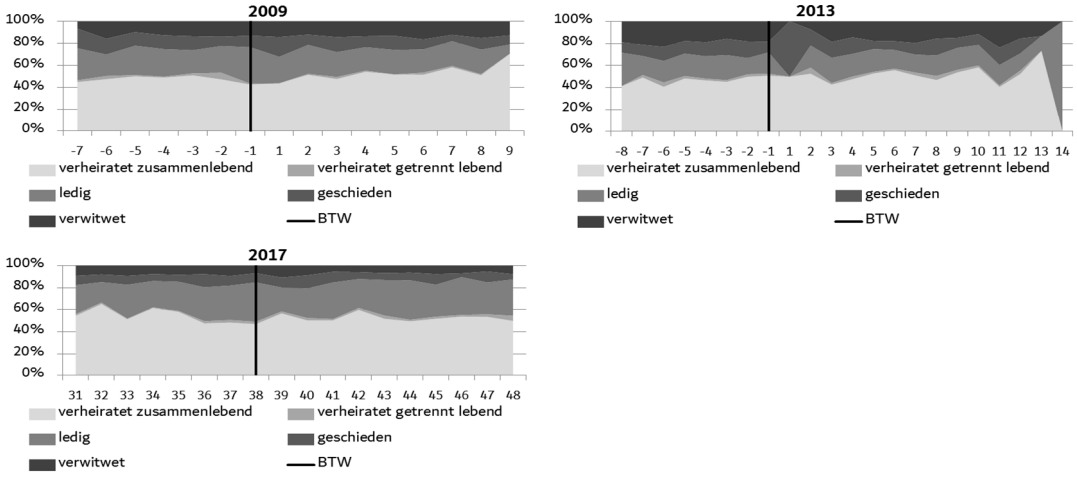
Berufliche Stellung



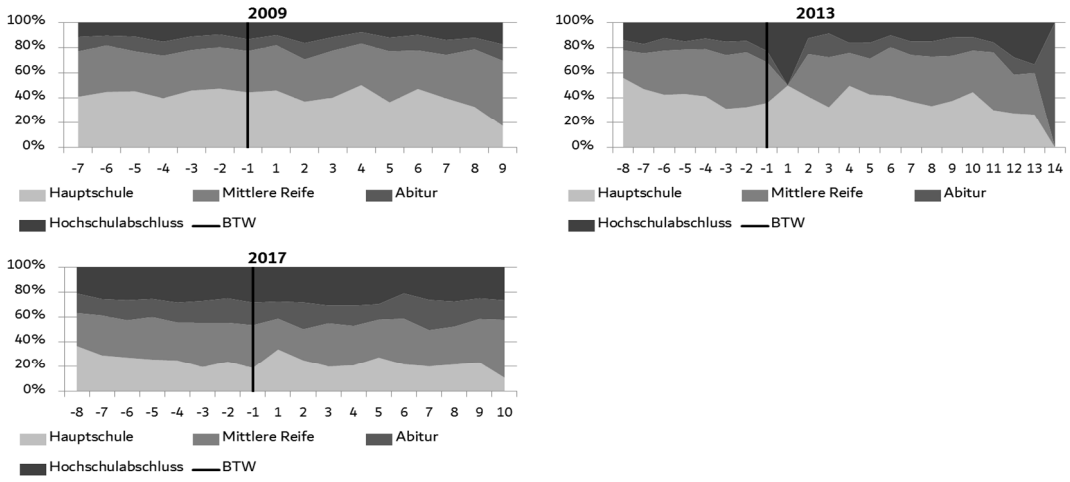




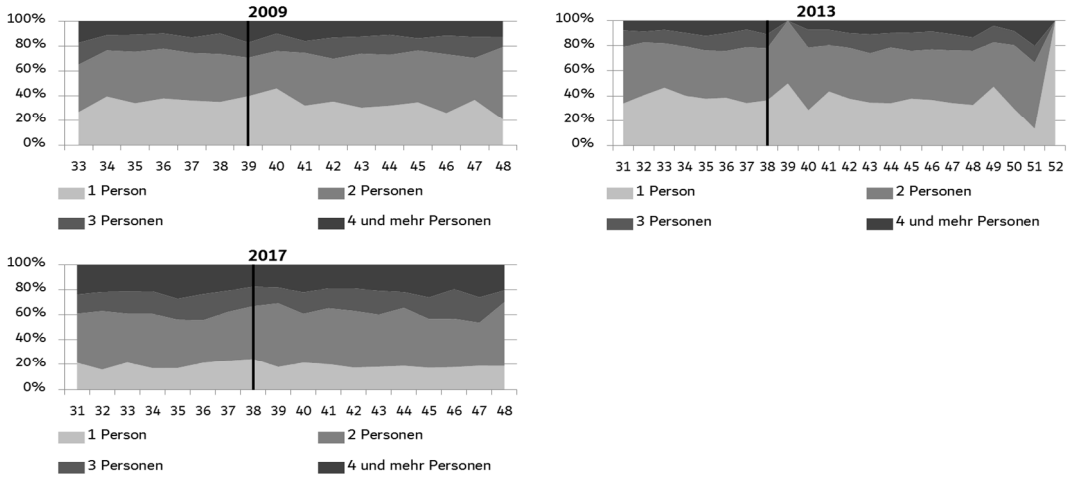
Familienstand



Bildung



Haushaltsgröße



Abbildungen 14: Feldverlauf der GLES-Querschnitte nach sozialstrukturellen Merkmalen in Kalenderwochen

2009 und 2017 sowie in den „restlichen“ Wochen der Befragung in 2013 zeigen sich lediglich leichte Schwankungen in den Verteilungen der jeweiligen Merkmalsausprägungen über die einzelnen Wochen der Feldzeit. Da hier keinerlei Systematik ausgemacht werden kann und sich – mit Ausnahme der eben für 2013 betrachteten Verzerrungen – keine groben Abweichungen zeigen, kann hier von einer annähernden Gleichverteilung der Ausprägungen der betrachteten Merkmale ausgegangen werden.

3.3 Weitere Indikatoren

Neben der quantitativen Darstellung der realisierten Interviews nach Woche und der sozialstrukturellen Verteilung bietet sich die Betrachtung verschiedener weiterer Indikatoren an, um den Feldverlauf zu beurteilen. Nachfolgend werden darunter insbesondere Maßzahlen verstanden, die die Entwicklungen im Feld in Abhängigkeit zu weiteren Faktoren (wie z.B. der Bruttostichprobe) setzen und dadurch Aussagen über die Qualität der Stichprobenrealisierung – und bei einer longitudinalen Entwicklung auch über den Fortschritt der Feldarbeit – zulassen. Die Betrachtung dieser Indikatoren bereits während der Feldzeit sind sehr bedeutend, um Schwierigkeiten im Feld möglichst frühzeitig zu erkennen und im besten Falle empiriebasiert Entscheidungen über Maßnahmen zur Gegensteuerung zu entwickeln.

Nachfolgend fokussieren wir uns auf sieben Indikatoren, um die Qualität der Feldarbeit zu betrachten. Vier der sieben Indikatoren sind bereits aus dem Kapitel 2.1.1, der Beschreibung des Teilnahmeverhaltens, bekannt (vgl. hierzu Bieber, Blumenberg, Blumenberg & Blohm, 2020):

(1) Ausschöpfungsquote,

(2) Kontaktrate,

(3) Kooperationsrate und

(4) Verweigerungsrate.

Um die Geschehnisse in der Feldarbeit zusätzlich beurteilen zu können, eigenen sich noch drei weitere Indikatoren (vgl. hierzu Bieber et al., 2020):

(1) Kontaktversuche, also der Anteil an der Bearbeitung von Adressen an der Bruttostichprobe,

(2) beendete Fälle, also der Anteil an durchgeführten Interviews an der Bruttostichprobe und

(3) final bearbeitete Adressen an der Bruttostichprobe, also Adressen, die in der Zukunft nicht mehr weiterbearbeitet werden, da entweder bereits ein Interview durchgeführt werden konnte oder ein finaler Ausfallgrund vorliegt, der keine weitere Adressbearbeitung zur Folge hat.[2]

Grundsätzlich können die Indikatoren für verschiedenste Einheiten analysiert werden: Deutschland insgesamt, Bundesländer, Sample-Point-Ebene, Interviewer-Ebene, u.a.. Im nachfolgenden Teil wird dies zunächst auf der Deutschlandebene (Kapitel 3.3.1) und der Bundesländerebene (Kapitel 3.3.2) realisiert. Kapitel 3.3.3 geht noch eine Ebene tiefer.

Ein besonderes Problem in der Feldarbeit, welches grundsätzlich auftreten kann, ist die mangelnde Performance von bestimmten Gebietskörperschaften. Dieses Problem ist in der Theorie einfach

[2] Von der Analyse ausgeschlossen sind nachbearbeitete Adressen. Nachbearbeitete Adressen sind Adressen, die zunächst als Verweigerungen gewertet wurden. Da es sich dabei jedoch um keine „harte“ Verweigerungen gehandelt hatte, wurden diese – um die Ausschöpfungsquote zu erhöhen - nachbearbeitet.

zu lösen: Sobald festgestellt wird, dass in bestimmten Gebieten die Feldbearbeitung problematisch ist, könnten gute Interviewer/innen dorthin entsandt werden, um den Bearbeitungsstatus vor Ort zu verbessern. Ein Problem bei der Betrachtung der oben beschriebenen Indikatoren ist jedoch, dass bei wachsender Zahl an betrachteten Ebenen die Unübersichtlichkeit steigt und es zunehmend schwierig ist, geographische Probleme zu erfassen, um darauf aufbauend oben dargestellte, angemessene Strategien zur Verbesserung der Feldarbeit entwickeln zu können. Die Darstellung der Indikatoren auf einer geographischen Karte stellt dagegen eine Lösung dar, diesem Problem zu begegnen. Bislang wird diese Methode in der Feldarbeit wenig eingesetzt. Möglichkeiten, wie dies zukünftig getan werden könnte, wurden von Bieber et al. (2020) für die Nachwahlbefragung 2017 entwickelt und sollen unter 3.3.3 vorgestellt werden.

3.3.1 Allgemeiner Überblick

Abbildung 15 gibt einen Überblick der sieben Indikatoren für die Vor- und Nachwahlbefragung des GLES-Querschnitts 2017. Beginnen wir mit der Ausschöpfungsquote: Sowohl in der Vor- als auch in der Nachwahlbefragung lag diese in der ersten Feldwoche bei ca. 20 Prozent, was nicht außergewöhnlich ist, da die Interviewer/innen zunächst die Haushalte aufsuchen, Interviewtermine ausmachen und bei Nichtantreffen einer Person im Haushalt weitere Kontaktversuche unternehmen müssen. In den folgenden drei Wochen steigt die Quote kontinuierlich an, woraufhin sie wieder etwas abfällt, was damit zusammenhängt, dass in diesen Wochen verstärkt Adressen, die final noch nicht zugeordnet wurden, eine finale Zuordnung erhalten und – je nach Zuteilung – dadurch die Ausschöpfungsrate wieder sinken kann.

Bezüglich der Verweigerungsquoten war zu erwarten, dass zunächst hohe Quoten zu verbuchen sind, da gerade in den ersten Wochen harte Verweigerungen, die eine finale Zuordnung zur Folge haben, verstärkt auftreten und dadurch hohe Verweigerungsraten erzeugt werden. Im Verlauf der Feldarbeit sanken diese 2017 in Vor- und Nachwahl auf unter 50 Prozent. Schließlich kann zu Beginn der Feldarbeit in beiden Studienteilen eine hohe Kontaktrate von über 90 Prozent festgestellt werden, die insbesondere bei der Vorwahlbefragung im Verlauf der Feldarbeit wiederum auf 80 Prozent sinkt.

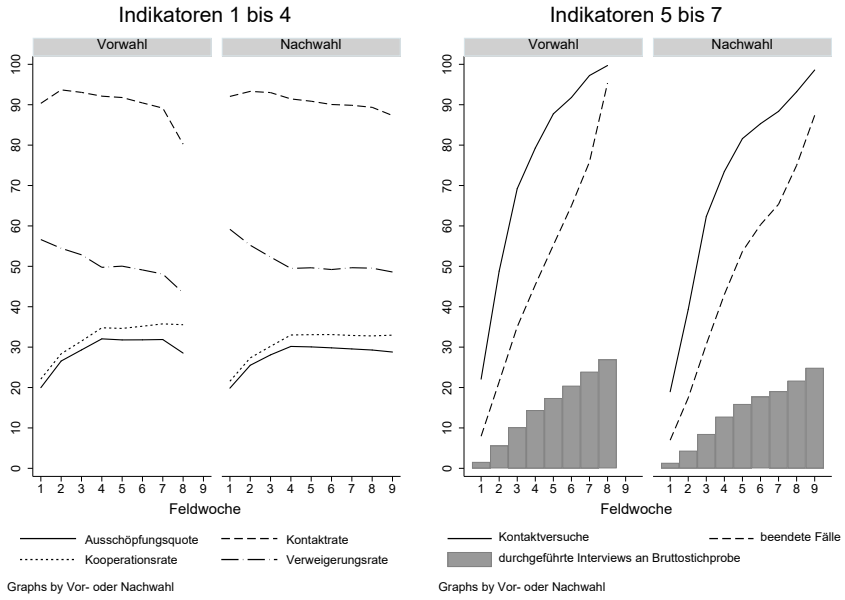


Abbildung 15: Indikatoren 1 bis 7 auf Deutschlandebene

Bezüglich der drei weiteren Indikatoren beschreiben folgende Verläufe einen idealen Feldverlauf:

1. Die durchgeführten Interviews an der Bruttostichprobe sollten von Woche zu Woche stetig steigen.
2. Die Kontaktversuche sollten bereits in Woche eins eine hohe Ausprägung aufweisen und von Woche zu Woche sehr steil ansteigen und in Woche drei möglichst Werte über 80 Prozent aufweisen. Dies würde bedeuten, dass 80 Prozent der Adressen in Woche drei bereits von einem Interviewer/ einer Interviewerin angefasst und bearbeitet wurden. Spätestens am Ende der Feldzeit, sollte die Kontaktrate annähernd 100 Prozent betragen. Das bedeutet, dass alle Adressen zumindest einmal bearbeitet wurden.
3. Schließlich ist zu erwarten, dass die beendeten Fälle ebenso von Woche zu Woche steigen und am Ende möglichst 100 Prozent betragen. Im Vergleich zu den Kontaktversuchen wächst diese Kurve jedoch langsamer an, da eine Zuteilung einer Adresse zu „bearbeitet" voraussetzt, dass entweder ein Interview durchgeführt wurde, eine harte Verweigerung stattfand (mögliche Ausfallgründe: Person spricht nicht genügend Deutsch, Person ist verzogen, Adresse existiert nicht, …) oder mindestens vier Kontaktversuche stattgefunden haben, um eine finale Zuordnung der Adresse vornehmen zu können. Auch in diesem Fall ist es wünschenswert, wenn die Rate am Ende der Feldzeit möglichst 100 Prozent beträgt.

Für die Vor- und Nachwahl 2017 zeigen die drei Indikatoren Folgendes: Die durchgeführten Interviews an der Bruttostichprobe zeigen, dass der Feldverlauf in der Vorwahl deutlich zügiger in Fahrt kam und höhere Realisierungsquoten zu verbuchen sind als in der Nachwahlerhebung. Ebenso ist zu sehen, dass in beiden Fällen mit einer hohen Kontaktfrequenz von 20 Prozent der Feldverlauf startete (Kontaktversuche), jedoch auch hier die Kurve bei der Vorwahlbefragung schneller und steiler angestiegen ist als dies in der Nachwahl der Fall ist. Das bedeutet, dass die Interviewer/innen ihre ersten Kontakte mit jeder Adresse in der Vorwahlbefragung deutlich schneller durchführten als in der Nachwahl. Die Verläufe zeigen jedoch auch, dass mit den meisten der Adressen zumindest ein Kontaktversuch durchgeführt wurde. Die Linie der bearbeiteten Fälle illustriert zudem, dass die finale Zuordnung der Fälle in der Vorwahlbefragung schneller und vollständiger vonstattenging. Auch wenn am Ende in der Vorwahl nur 95 Prozent der Adressen final zugeordnet wurden – was grundsätzlich optimiert werden kann – so zeigt sich bei der Nachwahlbefragung, dass dies nur bei knapp 90 Prozent der Fälle gelang. Dies bedeutet, dass am Feldende der Nachwahlbefragung noch immer bei zehn Prozent der Adressen die Möglichkeit bestanden hätte, ein Interview durchzuführen oder zumindest keine finale Zuordnung der Adresse stattfand. Auch dies zeigt, dass in der Nachwahlerhebung effizienter gearbeitet hätte werden können.

3.3.2 Bundesländern

Die sieben Indikatoren können auch auf der Ebene der Bundesländer berechnet werden. Sie sind in Abbildung 16 grafisch dargestellt. An dieser Stelle kann und soll keine detaillierte Analyse der Feldverläufe in den einzelnen Bundesländern erfolgen. Es soll genügen wichtige Besonderheiten und Probleme herauszuarbeiten:

(1) Variationen in den Verläufen: Es zeigt sich in den verschiedenen Darstellungen eine deutliche Variation in der Höhe und im Verlauf der einzelnen Indikatoren, sodass keinesfalls von einem bundesländerübergreifenden, harmonischen Feldverlauf zu sprechen ist. Fieldwork-Monitoring auf Ebene der Bundesländer ist daher angemessen und erforderlich.

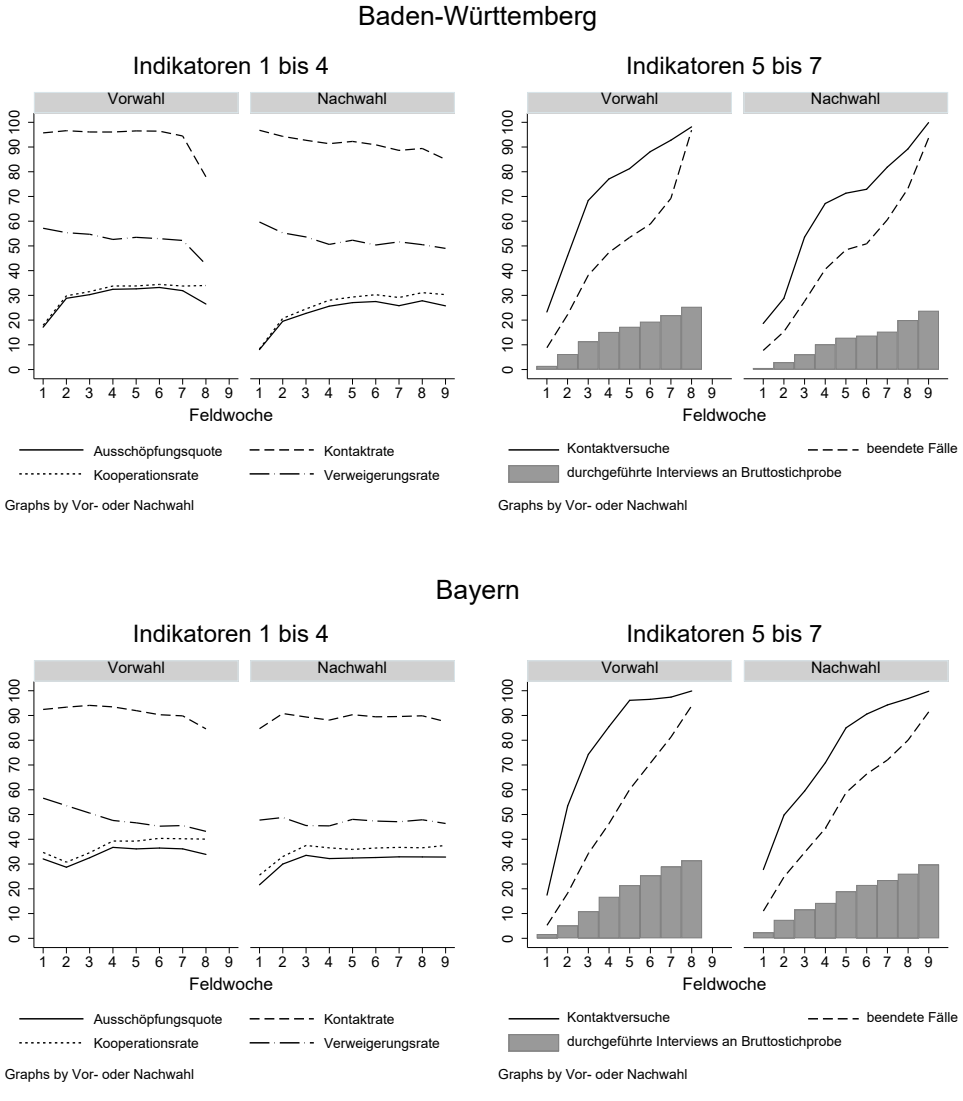
(2) Je kleiner die Bundeländer, desto instabiler die Verläufe: In den Stadtstaaten Berlin, Bremen und Hamburg und auch in dem bevölkerungsarmen Saarland beginnt die Feldarbeit in Teilen

verzögert (Saarland) und nimmt partiell sprunghafte Verläufe an (Bremen, Hamburg, Saarland), was zum Teil den niedrigen Fallzahlen geschuldet ist. Beispielsweise waren nur jeweils 48 Adressen im Saarland und in Bremen im Feld, was die sprunghaften Verläufe erklärt.

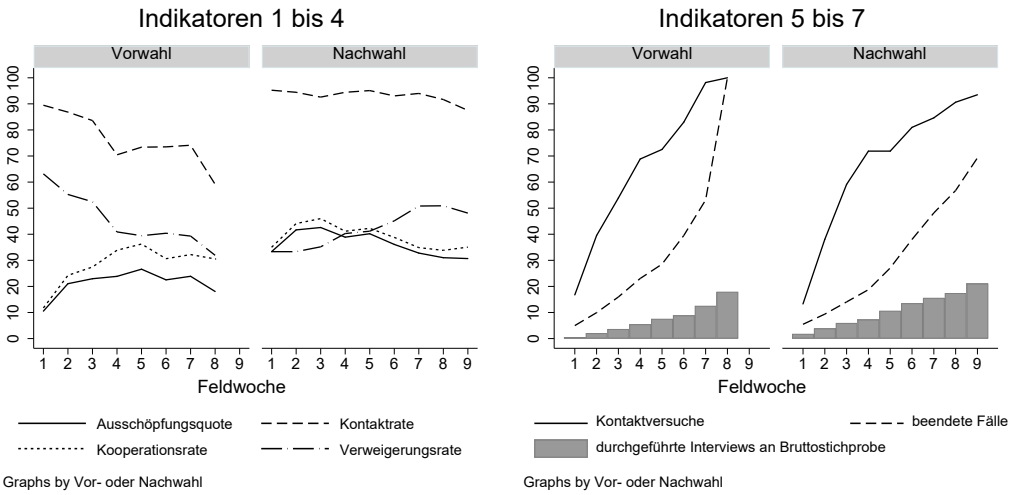
(3) Unterschiede zwischen Vor- und Nachwahl: In manchen Bundesländern unterscheiden sich die Verläufe in der Vor- und der Nachwahl deutlich (z.B. Hessen, Hamburg, Rheinland-Pfalz oder Sachsen-Anhalt). Dies zeigt uns, dass die Gefahr geographisch unterschiedliche verlaufender Feldarbeit berechtigt ist und stärkt das zukünftige Fieldwork-Monitoring darin, noch stärker auf homogene Feldverläufe in verschiedenen geographischen Einheiten zu achten.

(4) Mangel an finalisierten Adressen in bestimmten Bundesländern: Wie bereits die Ergebnisse für Deutschland insgesamt gezeigt haben, wurden in der Nachwahl viele Adressen nicht final bearbeitet. Besonders stark kann dies in Berlin, Hamburg, Sachsen und Sachsen-Anhalt beobachtet werden. Verhältnismäßig hohe Raten können dagegen in Bremen, Rheinland-Pfalz, Saarland, Schleswig-Holstein und Thüringen beobachtet werden.

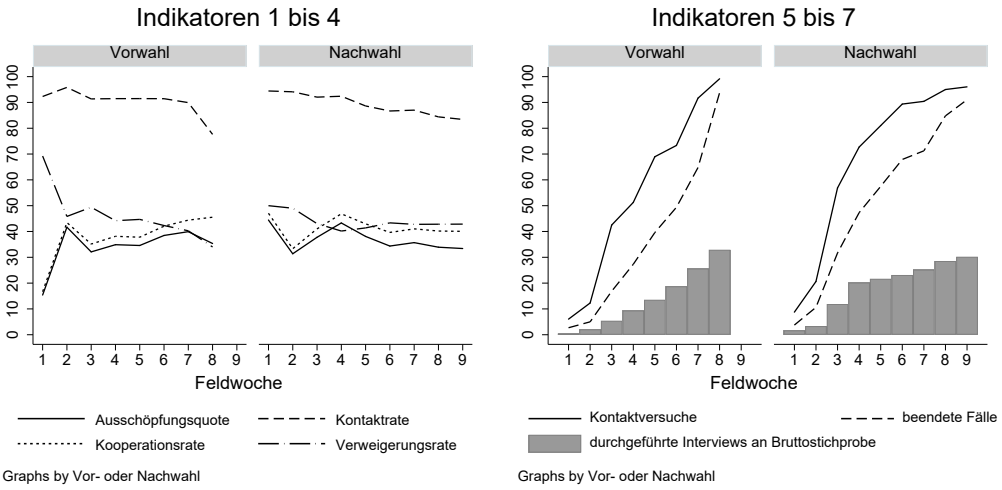
(5) Langsamer Feldstart in der Nachwahlbefragung: Insbesondere in der Nachwahlbefragung ist in manchen Bundesländern ein langsamer Feldstart zu beobachten. So finden in Baden-Württemberg, Bayern und Nordrhein-Westfalen in den ersten Wochen nach der Bundestagswahl verhältnismäßig wenig Kontaktversuche statt.



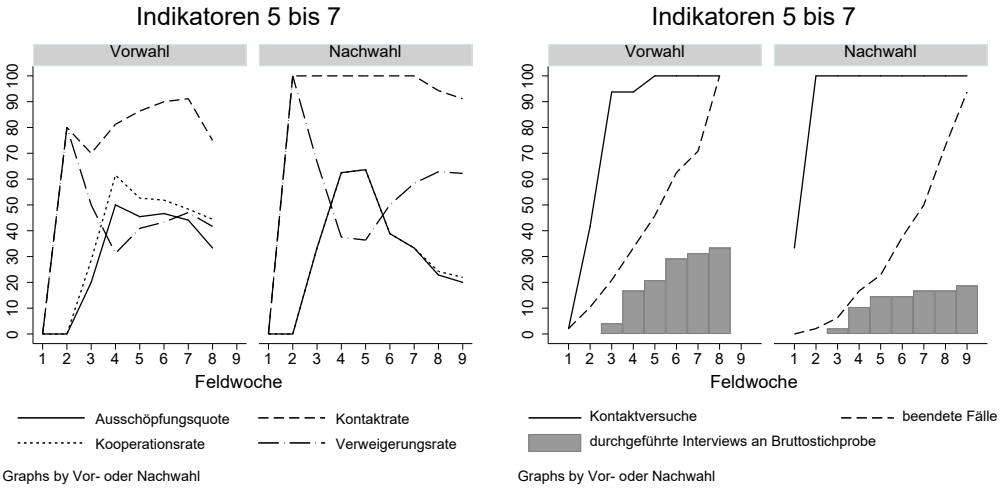
Berlin



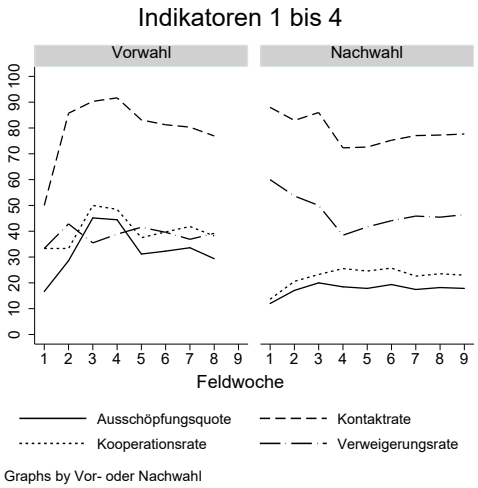
Brandenburg

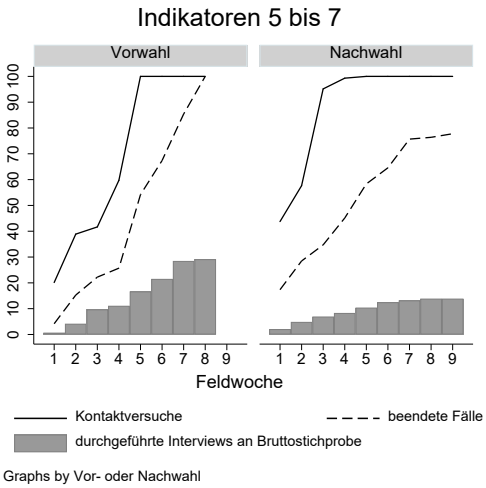


Bremen

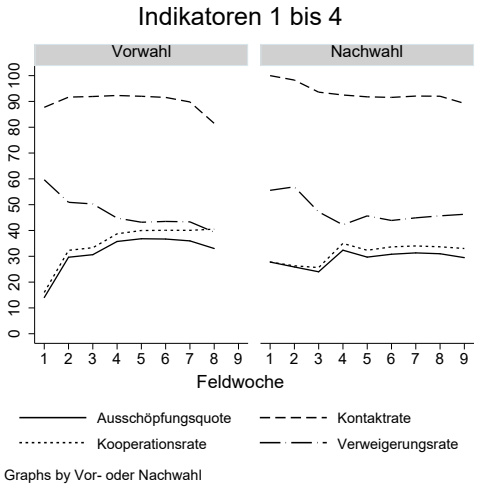


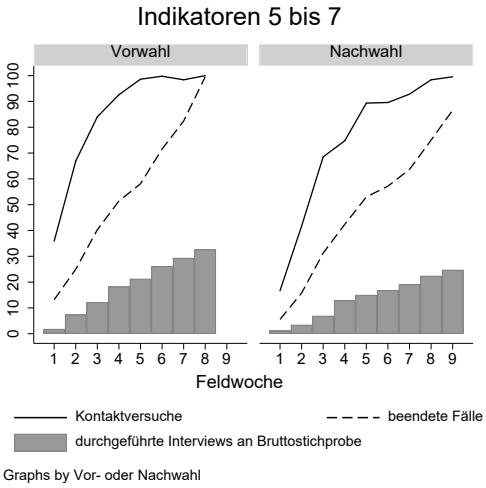
Hamburg



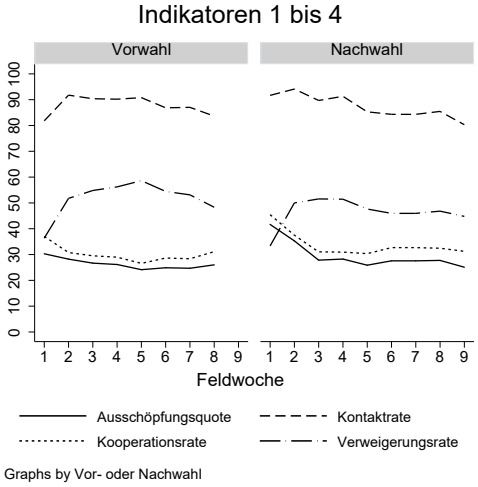


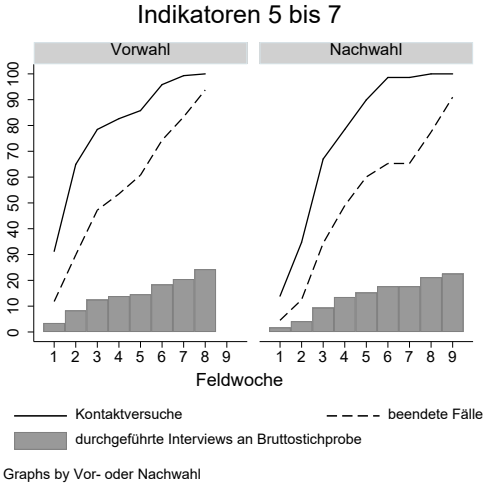
Hessen

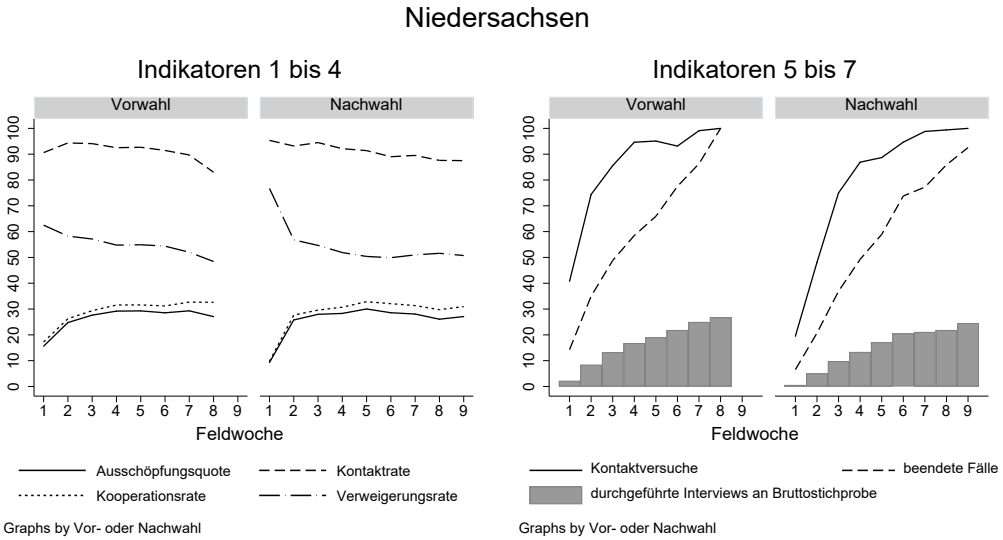


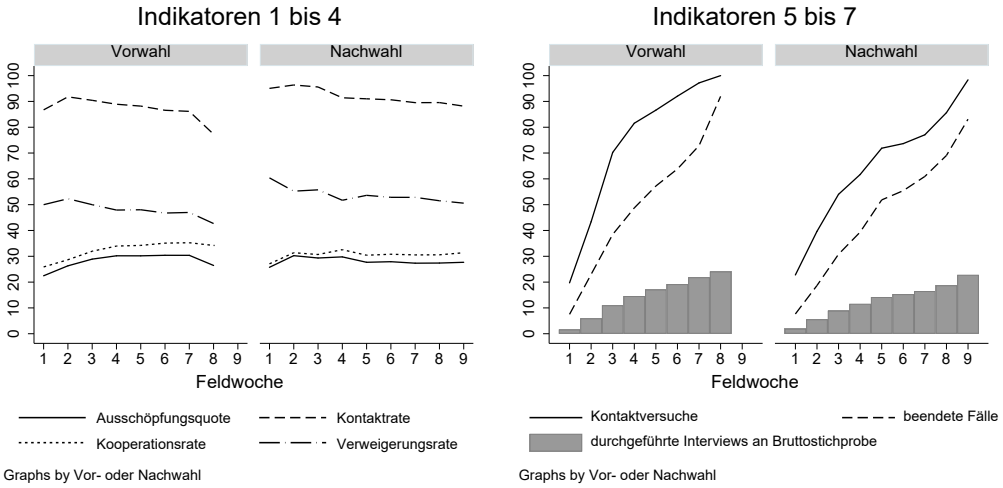


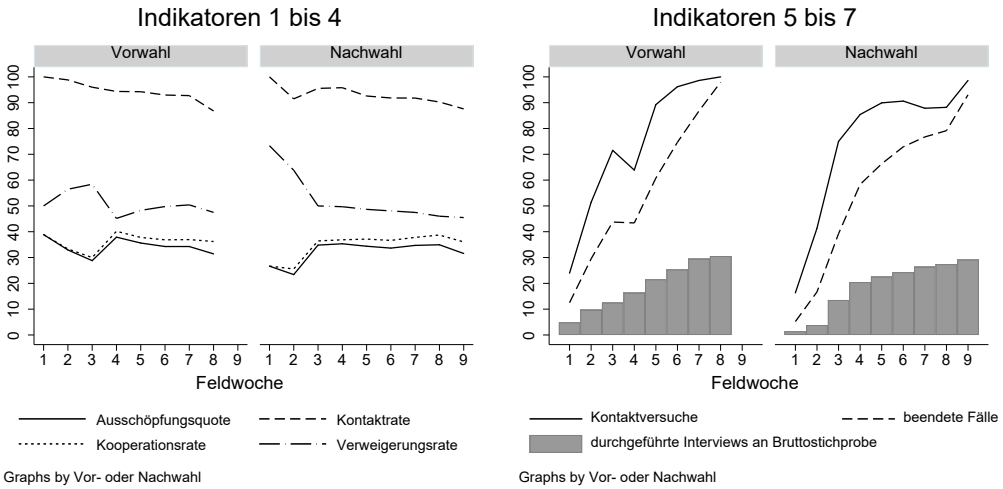
Mecklenburg-Vorpommern



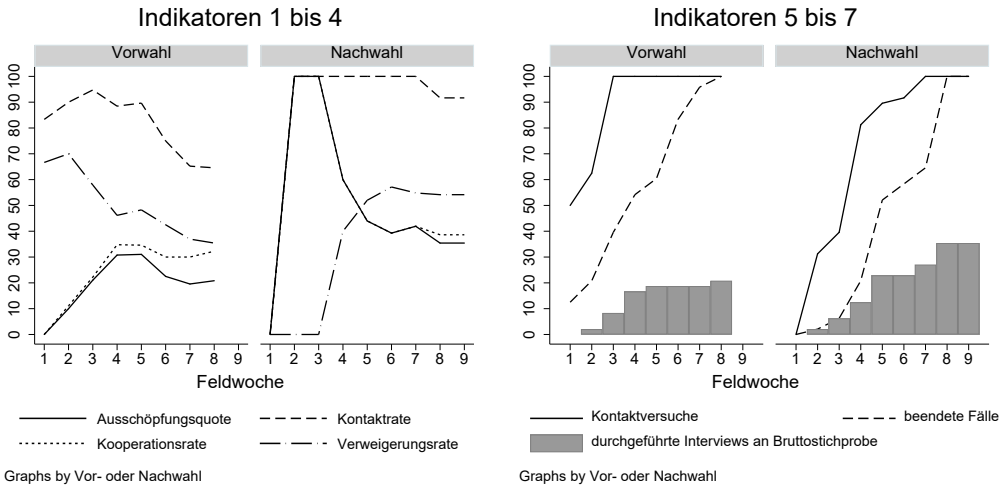



Niedersachsen
Indikatoren 1 bis 4
Indikatoren 5 bis 7
Vorwahl
Nachwahl
0 10 20 30 40 50 60 70 80 90 100
1 2 3 4 5 6 7 8 9
Feldwoche
Ausschöpfungsquote
Kooperationsrate
Kontaktrate
Verweigerungsrate
Kontaktversuche
beendete Fälle
durchgeführte Interviews an Bruttostichprobe
Graphs by Vor- oder Nachwahl

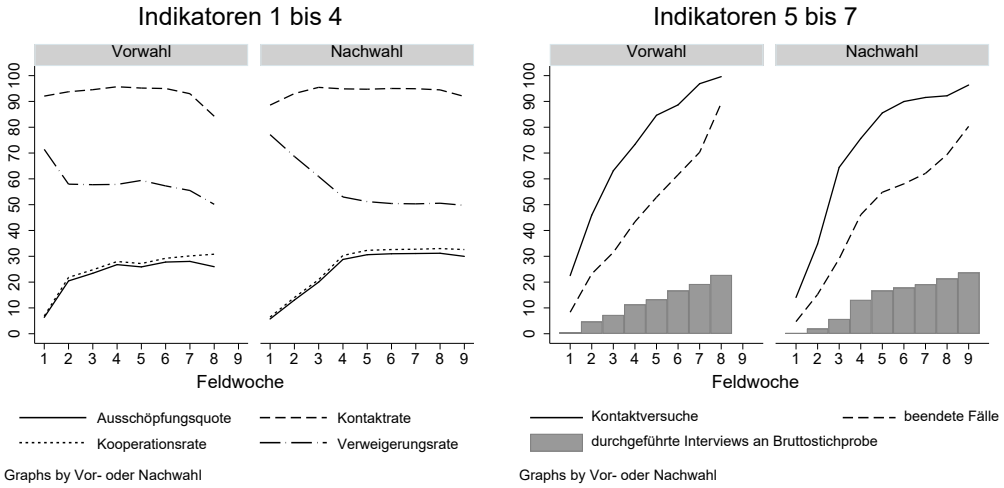

Nordrhein-Westfalen
Indikatoren 1 bis 4
Indikatoren 5 bis 7
Vorwahl
Nachwahl
0 10 20 30 40 50 60 70 80 90 100
1 2 3 4 5 6 7 8 9
Feldwoche
Ausschöpfungsquote
Kooperationsrate
Kontaktrate
Verweigerungsrate
Kontaktversuche
beendete Fälle
durchgeführte Interviews an Bruttostichprobe
Graphs by Vor- oder Nachwahl


Rheinland-Pfalz
Indikatoren 1 bis 4
Indikatoren 5 bis 7
Vorwahl
Nachwahl
0 10 20 30 40 50 60 70 80 90 100
1 2 3 4 5 6 7 8 9
Feldwoche
Ausschöpfungsquote
Kooperationsrate
Kontaktrate
Verweigerungsrate
Kontaktversuche
beendete Fälle
durchgeführte Interviews an Bruttostichprobe
Graphs by Vor- oder Nachwahl

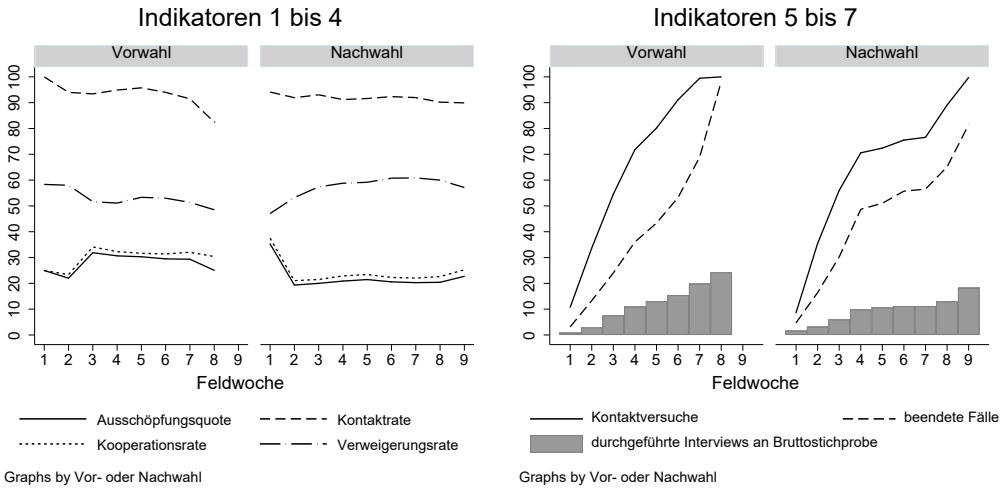
Saarland



Sachsen



Sachsen-Anhalt



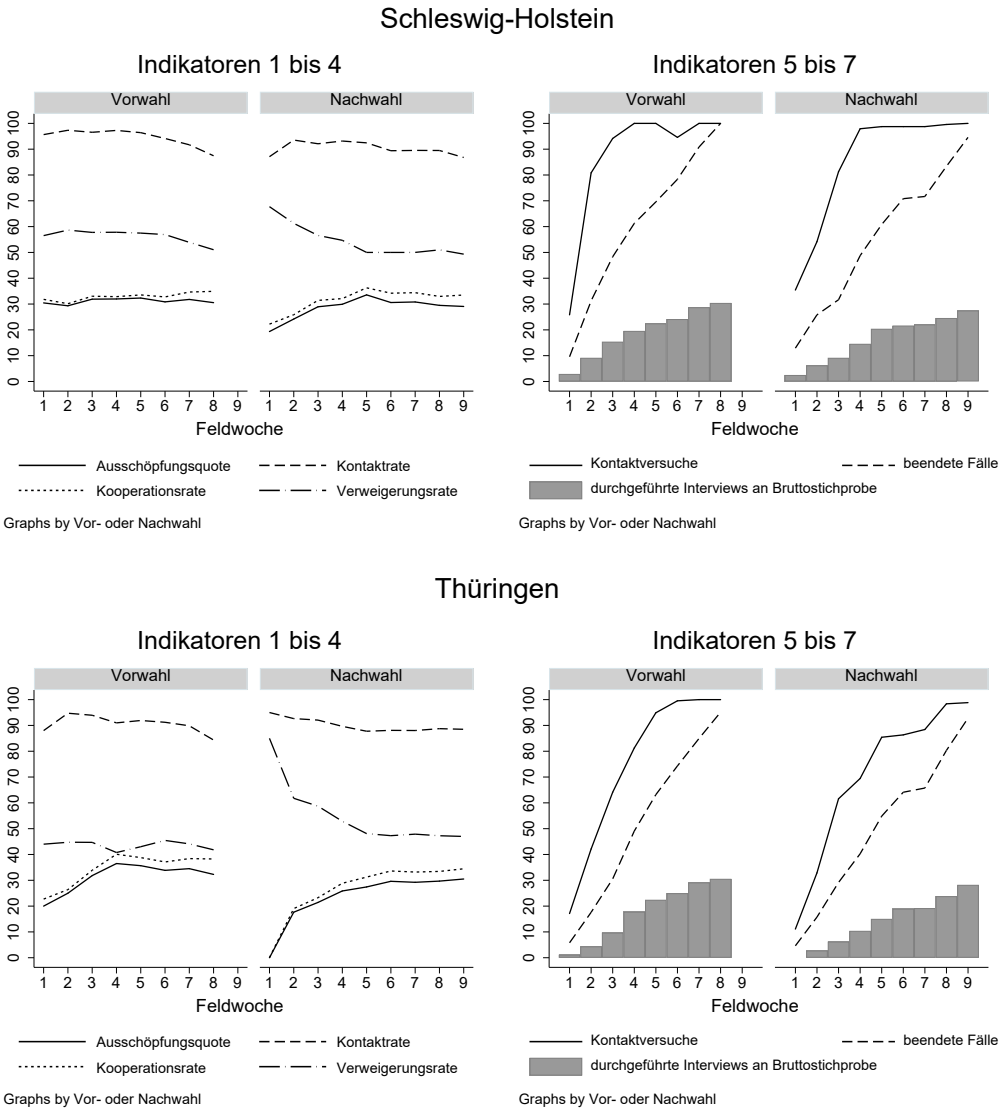


Abbildung 16: Indikatoren nach Vor- und Nachwahl 2017 in den verschiedenen Bundesländern

Zusammenfassend zeigen diese Darstellungen auf Bundeslandebene, dass bei der Feldarbeit 2017 zum Teil deutliche Unterschiede in den einzelnen Bundesländern beobachtet werden können. Diese Unterschiede sind gute Indizien, um problematische Feldverläufe zu identifizieren. Problem jedoch ist, dass auch die Bundeslandebene relativ grobkörnig ist, um daraus Maßnahmen abzuleiten, welche die Feldarbeit verbessern. Eine geographische Visualisierung, wie sie im nachfolgenden Teil dargestellt wird, kann helfen, Probleme auf geographischer Ebene zu lösen.

3.3.3 Geographische Darstellung

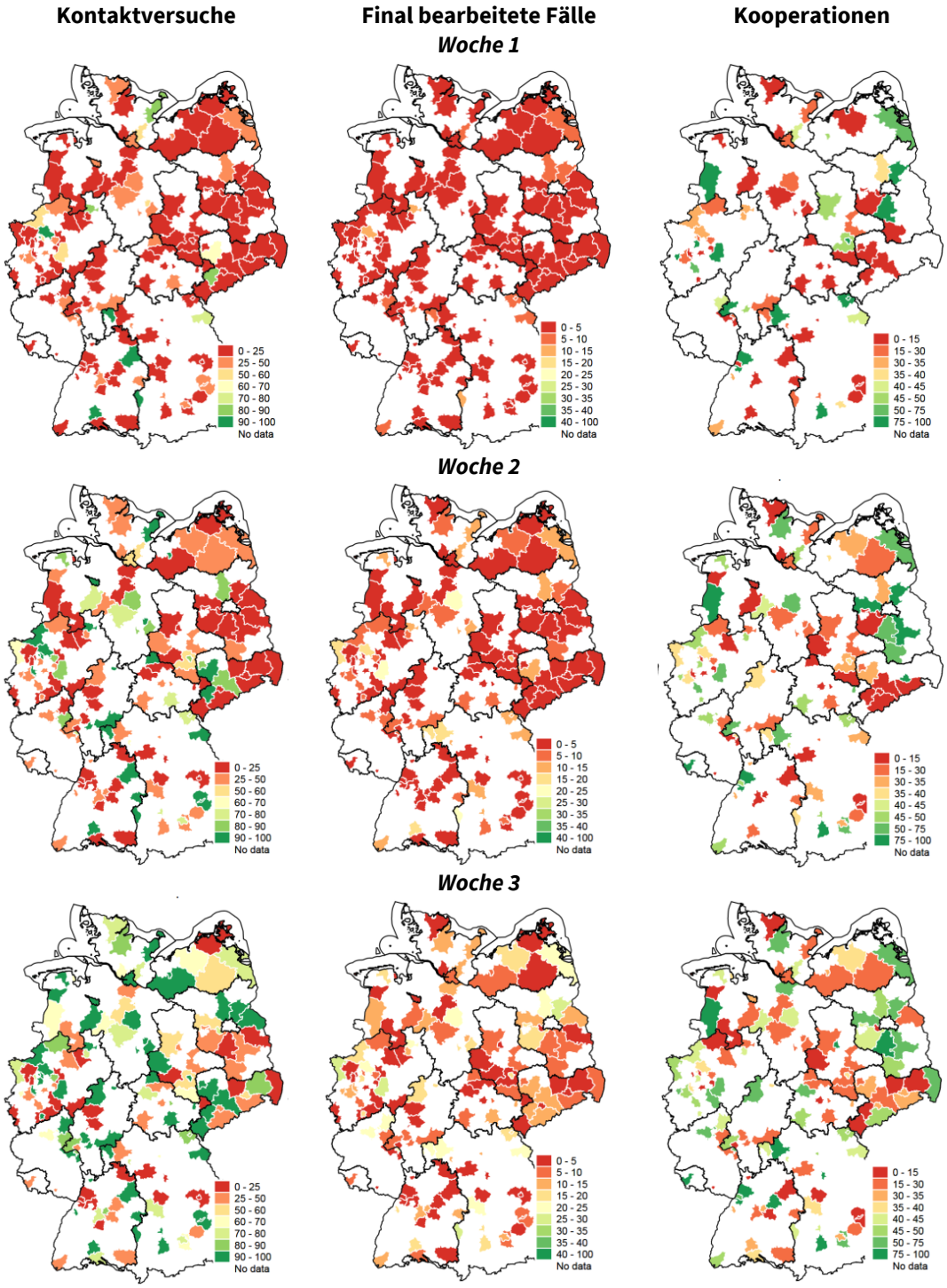
Wie bereits erwähnt haben Bieber et al. (2020) in einer Publikation die Indikatoren Kontaktversuch (Anteil an Kontaktversuchen an der Bruttostichprobe), final bearbeitete Fälle (Anteil an final bearbeitete Fälle an Bruttostichprobe) und Kooperationsrate (Anteil der Interviews an Interviews und Verweigerungen) auf geographischer Ebene visualisiert. Hintergrund bilden Probleme im Fieldwork-Monitoring bei der frühzeitigen Erkennung von gut und weniger gut arbeitenden geographischen Einheiten, um bei Feldproblemen angemessene Interventionsmaßnahmen einleiten zu können.

Hierzu wurden die Raten auf der NUTS3-Ebene berechnet und entlang einer Farbskala visualisiert. Es wurde die gängige Farbskala in Analogie zum Ampel-System verwendet, die von rot bis grün reicht. Auf der Karte bedeutet rot, dass die geographische Einheit zum aktuellen Zeitpunkt nicht oder mangelhaft bearbeitet ist, orange bedeutet eine mittelmäßige Bearbeitung und grün einen guten Bearbeitungsstatus.

Die nachfolgenden Grafiken zeigen für die Nachwahlerhebung 2017 – Woche für Woche – den Fortschritt des Bearbeitungsstatus nach Kontaktversuch (linkes Schaubild), abgeschlossene Fälle (mittleres Schaubild) und Kooperation (rechtes Schaubild). Die Grafiken und Beschreibung der Grafiken sind dem Aufsatz von Bieber et al. (2020) entnommen.

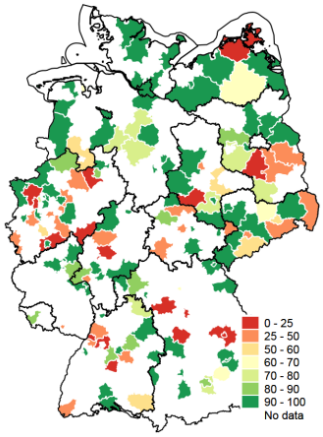
Die linke Grafik gibt Auskunft über die Kontaktversuche. Da die Interviewer/innen ihre Arbeit möglichst schnell aufnehmen und Kontakte zu den Befragten herstellen sollten, ist zu erwarten, dass sich die Karten relativ schnell von rot (bis zu 25% der Adressen wurden kontaktiert) zu grün (90 bis 100 Prozent der Adressen wurden kontaktiert) wechselt, was in zahlreichen Gebieten auch zu beobachten ist. Bei den mittleren Grafiken, die die abgeschlossenen Fälle darstellen, ist davon auszugehen, dass es etwas länger dauert, bis sich die Farbe von rot nach grün ändert, da vor dem tatsächlichen Interview häufig Kontaktversuche durchgeführt werden und sich die tatsächliche Durchführung von Interviews verzögert. Schließlich zeigt die rechte Karte das Verhältnis von erfolgreichen Interviews zu Interviewablehnungen plus Interviews. Die Veränderung der Farbgebung hierbei ist nicht zeitabhängig, was dazu führt, dass sich Gebiete auch von grün wieder zu rot ändern können.

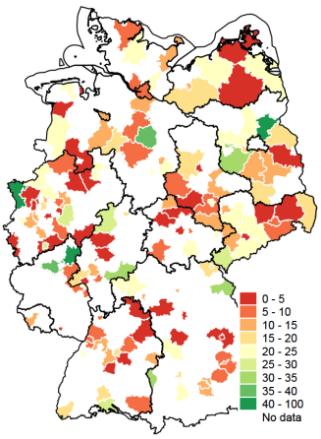
Dabei sind durchaus Gegenden zu entdecken, die auch nach mehreren Wochen noch immer einen problematischen Bearbeitungsstand aufweisen (z.B. Bundesländergrenzen im Ruhrgebiet oder zwischen Thüringen und Sachsen-Anhalt) und – mit einer Visualisierung während der Feldzeit – die Entsendung guter Interviewer/innen in problematische Gebiete eine Möglichkeit gewesen wäre, um den Bearbeitungsstand und somit regionale Unterschiede der Bearbeitung systematisch entgegenzuwirken. Analysedetails sind nachzulesen in Bieber et al. (2020).

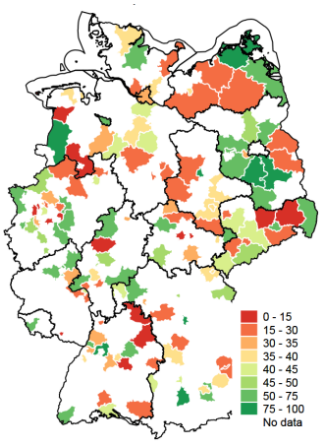

Kontaktversuche
Final bearbeitete Fälle
Kooperationen
Woche 1
Woche 2
Woche 3
0 - 25
25 - 50
50 - 60
60 - 70
70 - 80
80 - 90
90 - 100
No data
0 - 5
5 - 10
10 - 15
15 - 20
20 - 25
25 - 30
30 - 35
35 - 40
40 - 100
No data
0 - 15
15 - 30
30 - 35
35 - 40
40 - 45
45 - 50
50 - 75
75 - 100
No data

Kontaktversuche | **Final bearbeitete Fälle** | **Kooperationen**

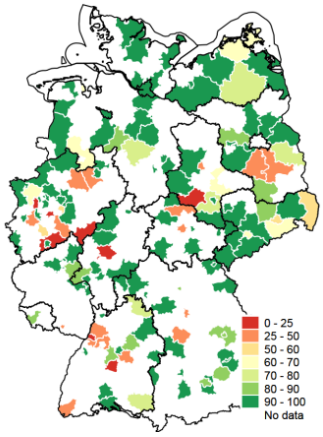
Woche 4

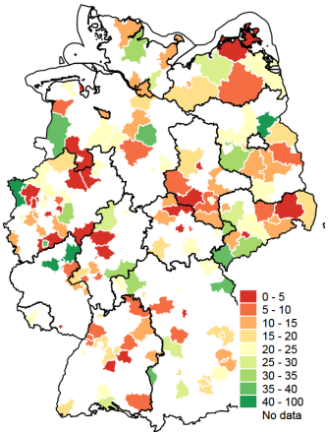


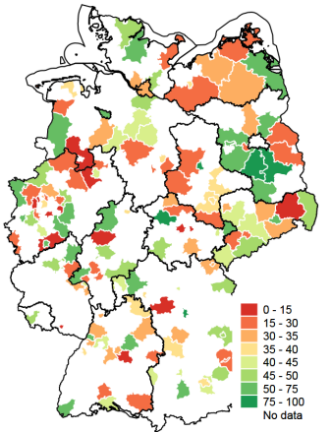




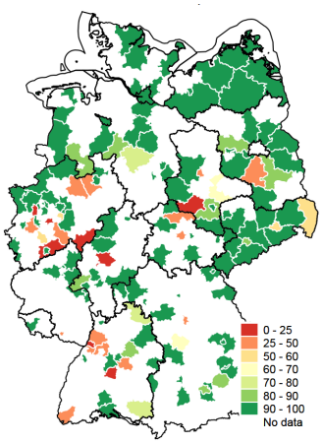
Woche 5

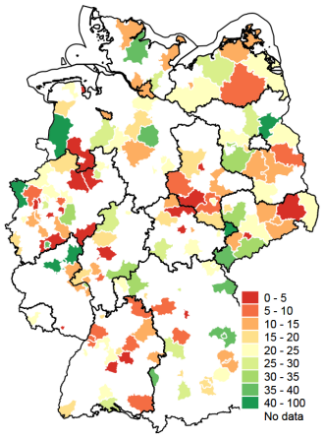


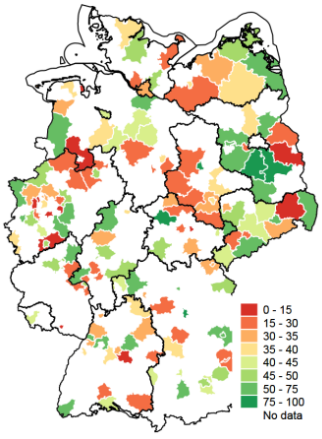




Woche 6







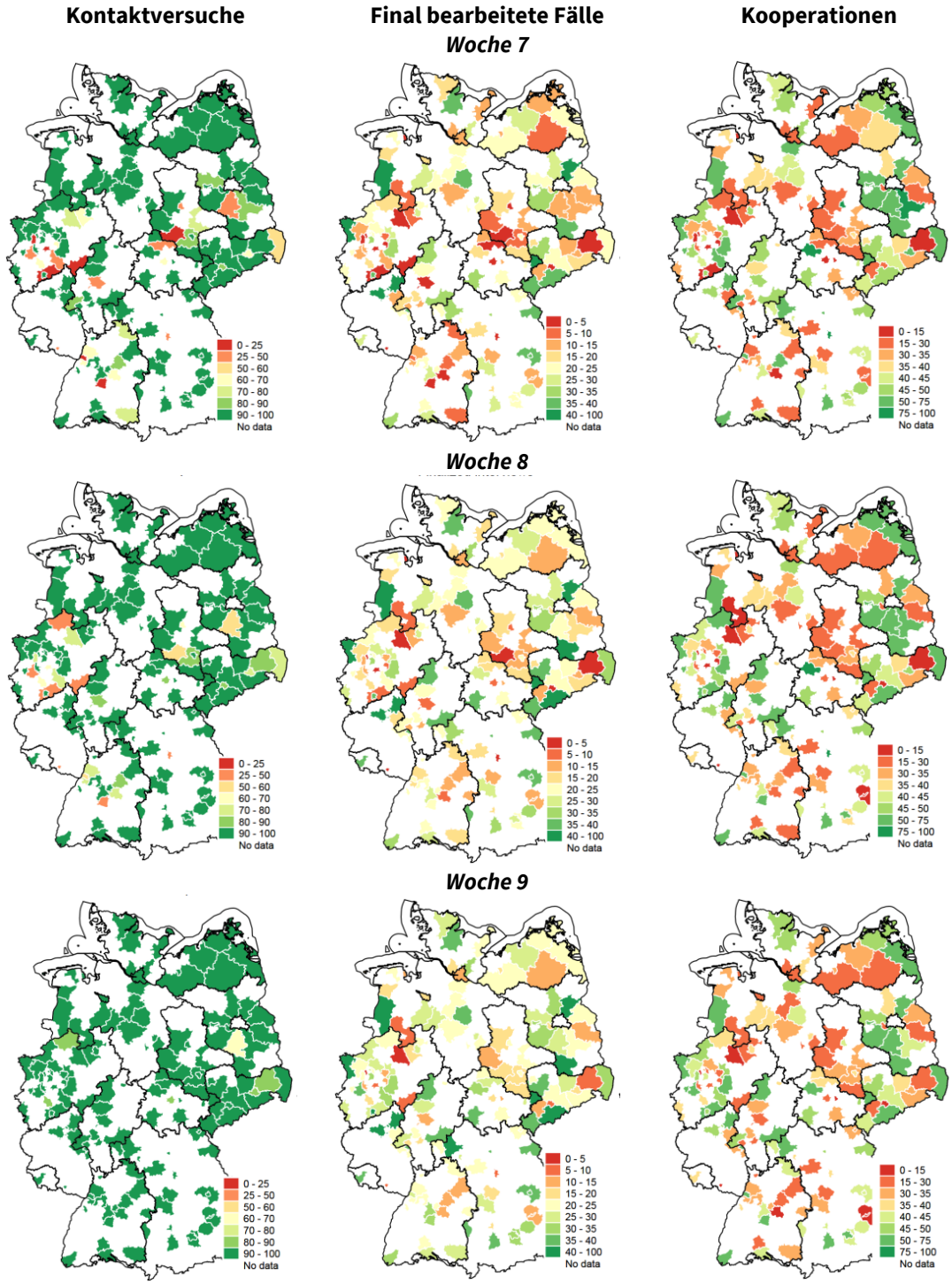


Abbildung 17: Geographische Darstellung der Feldarbeit

4 Zusammenfassung und Fazit

Sind persönlich-mündliche Interviews in Zeiten zunehmender Mobilisierung und Technisierung noch immer eine geeignete Methode für die Sozialforschung im Allgemeinen und die Wahlforschung im Besonderen, um qualitativ hochwertige Daten zu generieren? Obwohl diese einleitend gestellte Frage auch am Ende des vorliegenden Qualitätsberichts nicht eindeutig mit „ja“ oder „nein“ beantwortet werden kann, zeigt dieser Bericht sehr detailliert, wie die Datenqualität der GLES-Querschnitte grundsätzlich beschaffen ist und wie sie sich über drei Erhebungszeitpunkte – differenziert nach Vor- und Nachwahlbefragung – entwickelt hat.

Das Teilnahmeverhalten in den GLES-Querschnitten ist – insbesondere in Bezug auf den wichtigen Indikator „Ausschöpfung“ – mit dem gemessenen Verhalten in anderen Studien wie dem ALLBUS (2018) oder dem EVS vergleichbar (Christmann et al., 2019). Im longitudinalen Vergleich zeigt sich, dass – ebenso ähnlich wie in anderen Studien – die Kontaktversuche erfolgten, aber im Laufe der Zeit zu geringeren Kooperations- und Ausschöpfungsraten führten. Diese Abfälle waren insbesondere von 2009 zu 2013 zu beobachten. 2017 konnte das Niveau von 2013 gehalten bzw. leicht gesteigert werden. Ein Befund, der für die zukünftige Durchführung von persönlich-mündlichen Interviews besonders wichtig ist, ist die steigende Verweigerungsrate. Dieser Trend zeigt, dass die Interviewer/innen der GLES zunehmend Probleme haben, die ausgewählte Zielperson davon zu überzeugen, an der Umfrage teilzunehmen.

Doch der Bericht zeigt nicht nur eine sinkende Ausschöpfungsquote. Diese wäre nicht weiter problematisch, wenn die teilnehmenden Personen repräsentativ für die Grundgesamtheit stehen würden. Es zeigt sich jedoch auch, dass die Ausschöpfungsquote zwischen verschiedenen sozialstrukturellen Gruppen variiert. 2017 wurden insbesondere Personen zwischen 20 und 49 Jahre und über 70 Jahre unterrepräsentiert. Ebenso konnte beobachtet werden, dass Unterschiede der Ausschöpfungsquoten in verschiedenen Bundesländern auftraten. Dies zeigt, dass der Blick auf die Ausschöpfungsquote allein nicht genügt, sondern auch die Qualität der ausgeschöpften Fälle betrachtet werden muss. Die Ergebnisse legen somit nahe, dass in zukünftigen Erhebungen sowohl in der Planungsphase als auch während der Erhebung derartige Verzerrungen mitgedacht und Kontrollmechanismen entwickelt werden müssen, um zu versuchen die Datenqualität auch zukünftig erhalten zu können.

Um das Antwortverhalten der Befragten zu bewerten, wurde dieses in einem zweiten Schritt in Bezug zu Referenzstudien wie dem Mikrozensus oder dem tatsächlichen Wahlergebnis gesetzt. In den betrachteten Merkmalen Geschlecht, Alter, Bildung, berufliche Stellung, Familienstand und Haushaltsgröße zeigen sich nahezu überall Abweichungen von den Referenzwerten. Männer sind in den GLES-Querschnitten über- und Frauen unterrepräsentiert, wobei im Jahr 2017 eine deutliche Verbesserung zu erkennen ist. Hier wird das Geschlechterverhältnis adäquat abgebildet. Ein ähnliches Bild zeigt sich auch bei der Verteilung der Altersgruppen. Während 2013 jüngere Altersgruppen deutlich unter- und ältere überschätzt werden, zeigen sich 2009 und 2017 – mit kleineren Abweichungen – gute Abbildungen der tatsächlichen Verteilung von Altersgruppen. Eine deutliche Verschlechterung über die Erhebungsjahre hinweg zeigt sich bei der Bildung. Zwar gibt es auch 2009 und 2013 teils große Abweichungen vom jeweiligen Referenzwert, jedoch verschärft sich dieses Problem 2017 nochmals: Personen mit niedrigem Bildungsgrad werden deutlich unter-, solche mit hohem Bildungsgrad deutlich überschätzt. Weniger problematisch fällt die Betrachtung der beruflichen Stellung aus. Mit Ausnahme einer starken Überschätzung von Arbeiter/innen im Jahr 2013 zeigen sich ansonsten nur kleinere Abweichungen vom Mikrozensus. Beim Familienstand und der Haushaltsgröße spiegeln sich dagegen abermals Probleme persönlicher Interviews wider. Ledige Personen sowie Einpersonenhaushalte sind in allen Erhebungsjahren unter-, verhei-

ratete Personen sowie Zweipersonenhaushalte überschätzt. Zwar verbessert sich die Abbildung im GLES-Querschnitt 2017 ein wenig, da hier die Abweichungen geringer ausfallen – und sich zumindest der Anteil von Zweipersonenhaushalten nicht signifikant vom Referenzwert unterscheiden –, allerdings bleiben vor allem Einpersonenhaushalte und Singles schwerer erreichbar.

Auch wenn einige der betrachteten soziodemografischen Merkmale kaum bzw. keine signifikanten Abweichungen zu den Anteilswerten des Mikrozensus aufweisen, zeigen sich doch teils deutliche Verzerrungen in Merkmalen, die in der Planung zukünftiger Erhebungen berücksichtigt werden müssen: die Unterrepräsentation niedrig gebildeter Befragter sowie die geringere Erreichbarkeit lediger und/oder alleinlebender Personen.

Ebenso konnten Verzerrungen des erhobenen Wahlverhaltens im Vergleich zum tatsächlichen Wahlverhalten bei den Bundestagswahlen 2009, 2013 und 2017 beobachtet werden. Als besonders problematisch fällt die Wahlbeteiligung ins Gewicht: So ist zu erkennen, dass zu allen drei Befragungszeitpunkten die Wahlbeteiligung überschätzt wird. In der realisierten Stichprobe sind somit proportional gemessen zu viele Wähler/innen und zu wenige Nichtwähler/innen enthalten, was insbesondere die Nichtwahlforscher/innen vor inhaltliche sowie methodische Probleme stellt. Doch es ist nicht nur so, dass die Wahlteilnahme grundsätzlich überschätzt wird; zusätzlich ist eine Steigerung der Überschätzung von Wahljahr zu Wahljahr zu beobachten. Und auch die Wahlentscheidung konnte durch die GLES-Querschnitte nicht immer exakt erfasst werden. Ähnlich wie in anderen Wahlumfragen wurde tendenziell der Anteil an Grüne-Wähler/innen über- und der Anteil an AfD-Wähler/innen unterschätzt (Roth & Wüst, 2015, S. 311). Zwar ist es nicht Aufgabe der GLES-Querschnitte, das Wahlverhalten exakt zu erfassen, dennoch zeigen die Ergebnisse, dass die Werte – insbesondere in Bezug auf die Wahlbeteiligung – abweichen. Die sinkenden Anteile an Nichtwähler/innen deuten auf grundlegende Probleme hin, die direkte Auswirkungen auf die Möglichkeiten und Grenzen der Forschung haben. Es ist daher wichtig, diesen Punkt bei der zukünftigen Realisation der GLES-Querschnitte mitzudenken und geeignete Gegenmaßnahmen zu entwickeln.

Ebenso wurde der Feldverlauf unter die Lupe genommen. Dabei sind folgende vier Punkte aufgefallen:

(1) *Zwei Normalverteilungen statt einer Normalverteilung*: Wünschenswert ist ein Feldverlauf des GLES-Querschnitts, der die Form einer Normalverteilung über die gesamten 16 Wochen Erhebungsphase der Vor- und Nachwahl annimmt. Dieser Wunsch wurde leider nie Wirklichkeit. Vielmehr können Annäherungen an Normalverteilungen in der jeweiligen Feldphase für die Vor- und die Nachwahlerhebungen beobachtet werden, was jedoch auch auf die getrennte Stichprobenziehung und die damit verbundene Feldarbeit zurückzuführen ist.

(2) *Nachwahl performt schlechter als Vorwahl*: Ein zweiter, wichtiger Befund ist, dass der Feldverlauf der Vorwahl deutlich besser als der der Nachwahl einzustufen ist. So konnte in zwei der drei Nachwahlbefragungen ein schleppender Feldstart der Nachwahl beobachtet werden. Es scheint gar so, dass die Interviewer/innen nach dem harten Ende der Vorwahlerhebung (am Tag vor der Bundestagswahl ist der letztmögliche Termin, um bearbeitete Adressen zu finalisieren) zunächst ermüdet sind und sich eine Pause von der GLES-Studie zu gönnen scheinen. Aufgrund der neuen Stichprobe müssen die Interviewer/innen zudem zunächst Kontakt zu den Zielpersonen aufbauen, was die Arbeit ebenso verzögert. Daher ist es besonders wichtig, zukünftig auf den Feldstart der Nachwahl zu schauen und frühzeitig Maßnahmen zu entwickeln, die diese Ermüdung erst gar nicht aufkommen lassen, und somit einen unverzüglichen Feldstart realisieren. Schließlich ist die Zeit kurz vor und kurz nach der Wahl für eine Wahlstudie besonders wichtig, da das abgefragte Verhalten dem tatsächlichen Verhalten am ehesten entspricht und keine Verzerrungen durch zeitliche Faktoren (Wahlkampf, Koalitionsbildungsprozesse, etc.) entstehen können.

(3) Homogene Verteilung sozialstruktureller Merkmale über die Feldzeit: Bei der Verteilung für die Wahlforschung relevanter sozialstruktureller Merkmale über den Feldverlauf können keine größeren Verzerrungen ausgemacht werden. Bis auf größere Ungleichverteilungen im GLES-Querschnitt 2013 in der Woche nach der Bundestagswahl und am Ende der Feldzeit, die allerdings auf die sehr geringe Zahl an realisierten Interviews zu diesen Zeitpunkten zurückgeführt werden können, und kleineren Schwankungen in allen Erhebungsjahren sind keine groben Abweichungen auszumachen.

(4) *Kein geographisch, homogener Feldverlauf:* Schließlich zeigen die Darstellungen verschiedener Indikatoren für das Jahr 2017 auf unterschiedlichen geographischen Ebenen, dass kein geographisch, harmonischer Feldverlauf beobachtet werden kann. Es gibt Bundesländer und Gebiete, in welchen die Feldarbeit gut absolviert wird und es gibt Bundesländer und Gebiete, die von Anfang an schlecht performen. Mit dem entwickelten Instrument von Bieber et al. (2020) kann im Rahmen des Fieldwork-Monitorings 2021 die geographische Performance beobachtet werden, um frühzeitig problematische Gegenden zu identifizieren, die dann mit Hochdruck bearbeitet werden können. Dies ist ein erster Ansatz, um dem Problem zu begegnen, wohlwissend, dass persönlich-mündliche Umfragen deutlich träger und unflexibler sind als Online- oder Telefonumfragen.

Zusammenfassend kann die Datenqualität der GLES-Querschnitte 2009, 2013 und 2017 als „gut“ bezeichnet werden, wenn auch gewisse Probleme hinsichtlich Teilnahmeverhalten (sinkende Ausschöpfungen/Kooperationsraten und steigende Verweigerungen, Unterschiede in Ausschöpfungen verschiedener sozialstruktureller Gruppierungen) und Antwortverhalten (insbesondere in Bezug auf Wahlbeteiligung) beobachtet werden konnten. Diese sind jedoch allesamt vergleichbar mit deren in anderen sozialwissenschaftlichen Face-to-Face-Umfragen. Mit diesem Bericht verfügen wir nun über ein umfangreiches Wissen von möglichen Problemen und Schwachpunkten von persönlich-mündlichen Erhebungen im Allgemeinen und von Wahlumfragen im Besonderen, die zukünftig bei der Planung und Umsetzung verschiedener Erhebungen berücksichtigen werden, um dadurch die Qualität langfristig zu erhalten und im besten Fall noch zu verbessern.

5 Literaturverzeichnis

AAPOR (2016). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. American Association for Public Opinion Research. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf. Zugegriffen: 16.09.2020.

ALLBUS (2019). ALLBUS 2018: Studien-Nr. 5270. GESIS - Leibniz-Institut für Sozialwissenschaften. www.gesis.org/allbus/inhalte-suche/studienprofile-1980-bis-2018/2018. Zugegriffen: 16.09.2020.

Bieber, I., Blumenberg, J. N., Blumenberg, M. S., & Blohm, M. (2020). Using Geospatial Data to Monitor and Optimize Face-to-Face Fieldwork. *Survey Methods: Insights from the Field, Special Issue: Fieldword Monitoring Strategies for Interviewer-Administered Surveys*, https://doi.org/10.13094/SMIF-2020-00005.

Bieber, I., & Bytzek, E. (2013). Herausforderungen und Perspektiven der empirischen Wahlforschung in Deutschland am Beispiel der German Longitudinal Election Study (GLES). *Analyse & Kritik*, *35*(2), 341–370.

Bieber, I., Roßteutscher, S., & Scherer, P. (2018). Die Metamorphosen der AfD-Wählerschaft: Von einer euroskeptischen Protestpartei zu einer (r)echten Alternative? *Politische Vierteljahresschrift*, *59*(3), 433–461.

Blumenberg, M. S., & Adewuyi, D. (2017). Feldverlauf. *In J. Roßmann, M.S. Blumenberg, T. Gummer (Hrsg.), Bericht zur Datenqualität der GLES 2013, GESIS Papers, 2017/13* (S. 25–43). Köln.

Blumenberg, M. S., Roßmann, J., & Gummer, T. (2013). *Bericht zur Datenqualität der GLES 2009.* GESIS - Technical Report 14. Mannheim.

Bortz, J., & Schuster, C. (2006). *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer.

Böth, K., & Kobold, K. (2013). Endgültiges Ergebnis der Wahl zum 18. Deutschen Bundestag am 22. September 2013. *WISTA - Wirtschaft und Statistik - Statistisches Bundesamt (Destatis)*, 845–861.

Brick, J. M., & Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 36–59.

Bundeswahlleiter (2018). Ergebnisse der repräsentativen Wahlstatistik: Tabellen zur Weiterverwendung: Zeitreihe seit 1953: Zweitstimmen nach Geschlecht und Altersgruppen (CSV-Datei). https://www.bundeswahlleiter.de/bundestagswahlen/2017/ergebnisse/repraesentative-wahlstatistik.html. Zugegriffen: 16.09.2020.

Christmann, P., Gummer, T., Hähnel, S., & Wolf, C. (2019). *Does the mode matter? An experimental comparison of survey responses between face-to-face and mixed-mode surveys.* ESRA Conference, Zagreb (17.7.2019).

Destatis (2019). Was ist der Mikrozensus? https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Haushalte-Familien/Methoden/mikrozensus.html. Zugegriffen: 16.09.2020.

FDZ der statistischen Ämter des Bundes und der Länder (2009). *Mikrozensus 2009*. eigene Berechnungen.

FDZ der statistischen Ämter des Bundes und der Länder (2013). *Mikrozensus 2013*. eigene Berechnungen.

Gabler, S., & Ganninger, M. (2010). Gewichtung. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 143–164). Wiesbaden: VS, Verlag für Sozialwissenschaften.

GESIS (2019). FDZ Wahlen. https://www.gesis.org/wahlen/wahlen-home. Zugegriffen: 16.09.2020.

GESIS (2020). GLES-Design. https://gles.eu/gles/das-gles-design/. Zugegriffen: 16.09.2020.

Gisart, B. (2009). Endgültiges Ergebnis der Wahl zum 17. Deutschen Bundestag am 27. September 2009. *WISTA - Wirtschaft und Statistik - Statistisches Bundesamt (Destatis)*, *11*, 1063–1079.

Groves, R. M., & Couper, M. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. M., & Lyberg, L. (2011). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065.

Hoover, E. M. (1936). The Measurement of Industrial Localization. *Review of Economics and Statistics*, *18*(4), 162–171.

Hugi, S. (2014). *Verzerrungen von selbstberichteten politischen Partizipationsangaben. Eine Validierungsstudie zu Abdeckungs-, Nonresponse- und Overreporting-Fehlern in der Schweizer Umfrageforschung*. Bern: Masterarbeit. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiI8aDDjevrAhWxM-wKHQRVDrwQF-jAAegQIAhAB&url=http%3A%2F%2Fsimonhugi.ch%2F%3Fddownload%3D141&usg=AOvVaw0ym1Arwi89mKwHs7erwpD6. Zugegriffen: 16.09.2020.

Kreuter, F. (2013). Facing the Nonresponse Challenge. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 23–35.

Kroh, M., & Käppner, K. (2016). Die Wirkung der Wahlbeteiligung auf das politische Interesse von Erstwählern. In H. Schoen & B. Weßels (Hrsg.), *Wahlen und Wähler: Analysen aus Anlass der Bundestagswahl 2013* (S. 371–397). Wiesbaden: Springer VS.

Leeuw, E. D. de, & Berzelak, N. (2016). Survey mode or survey modes? In C. Wolf, D. Joye, T. W. Smith, & Y.-c. Fu (Hrsg.), *The SAGE Handbook of Survey Methodology* (S. 142–156). London: Sage Publications.

Proner, H. (2011). *Ist keine Antwort auch eine Antwort? Die Teilnahme an politischen Umfragen*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wagner, A., Scherer, P., Bytzek, E., Bieber, I. (2019). *Vor- und Nachwahl-Querschnitt 2009 (Kumulation) (GLES 2009), Studienbeschreibung*. Mannheim. https://search.gesis.org/research_data/ZA5302. Zugegriffen: 16.09.2020.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wolf, C., Wagner, A., Giebler, H., Bieber, I., Scherer, P. (2019). *Vor- und Nachwahl-Querschnitt (Kumulation) (GLES 2013), Studienbeschreibung*. Mannheim. Zugegriffen: 16.09.2020.

Roßmann, J., Blumenberg, M. S., & Gummer, T. (2017). *Bericht zur Datenqualität der GLES 2013.* GESIS Papers, 2017/13. Köln.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., Wolf, C., Bieber, I., Stövsand, L.C., Dietz, M., Scherer, P.; Wagner, A.; Melcher, R.; Giebler, H. (2019). *Vor- und Nachwahl-Querschnitt (Kumulation) (GLES 2017), Studienbeschreibung*. Mannheim. https://search.gesis.org/research_data/ZA6802. Zugegriffen: 16.09.2020.

Roth, D., & Wüst, A. M. (2015). Missgeschick oder Trend? Zur Prognosetauglichkeit von Wahlumfragen. *Journal for Comparative Government and European Policy*, *13*(2), 298–323.

Schmitt-Beck, R., Rattinger, H., Roßteutscher, S., & Weßels, B. (2010). Die deutsche Wahlforschung und die German Longitudinal Election Study (GLES). In F. Faulbaum & C. Wolf (Eds.), *Gesellschaftliche Entwicklungen im Spiegel der empirischen Sozialforschung* (S. 141–172). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Opladen: Leske + Budrich.

Schumann, S. (2019). *Repräsentative Umfrage: Praxisorientierte Einführung in empirische Methoden und statistische Analyseverfahren*. Berlin: De Gruyter.

Sciarini, P., & Goldberg, A. C. (2015). Lost on the Way: Nonresponse and its Influence on Turnout Bias in Postelection Surveys. *International Journal of Public Opinion Research*, *17*, 110-137.

Sixtus, F., Slupina, M., Sütterlin, S., Amberger, J., & Klingholz, R. (2019). *Teilhabeatlas Deutschland: Ungleichwertige Lebensverhältnisse und wie die Menschen sie wahrnehmen*. Berlin. https://www.berlin-institut.org/studien-analysen/detail/teilhabeatlas-deutschland. Zugegriffen: 16.09.2020.

Stadtmüller, S., Silber, H., Daikeler, J., Martin, S., Sand, M., Schmich, P., Schröder, J., Struminskaya, B., Weyandt, K. W., & Zabal, A. (2019). *Adaptation of the AAPOR Final Disposition Codes for the German Survey Context* (GESIS – Survey Guidelines). Mannheim. https://www.gesis.org/fileadmin/upload/SDMwiki/2019_ResponseRates_Stadtmueller_1.pdf. Zugegriffen: 16.09.2020.

Stemmer, B. (2017). Endgültiges Ergebnis der Wahl zum 19. Deutschen Bundestag am 24. September 2017. *WISTA - Wirtschaft und Statistik - Statistisches Bundesamt (Destatis)*, 74–94.

Voogt, R. J. J., & Saris, W. E. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of Official Statistics*, *21*(3), 367–387.